UNIVERSIDADE FEDERAL DE SANTA CATARINA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Cláudio Pereira Flores

ELICITAÇÃO DO CONHECIMENTO TÁCITO DE DISTRIBUIÇÕES CONTÍNUAS DE PROBABILIDADE

Dissertação submetida à Universidade Federal de Santa Catarina como parte dos requisitos para obtenção do grau de Mestre em Ciência da Computação.

Prof. Paulo José de Freitas Filho, Dr.
(Orientador)

Profª. Silvia Modesto Nassar, Dra.
(Co-orientadora)

Florianópolis, agosto de 2008.

ELICITAÇÃO DO CONHECIMENTO TÁCITO DE DISTRIBUIÇÕES CONTÍNUAS DE PROBABILIDADE

Cláudio Pereira Flores

Esta Dissertação foi julgada adequada para a obtenção do título de Mestre em Ciência da Computação, área de concentração Sistemas de Conhecimento e aprovada em sua forma final pelo Programa de Pós-Graduação em Ciência da Computação.

_____

Prof. Frank Augusto Siqueira, Dr.
(Coordenador do Curso)

Banca Examinadora

_____

Prof. Paulo José de Freitas Filho, Dr.
(Orientador)

_____

Profª. Silvia Modesto Nassar, Dra.
(Co-orientadora)

_____

Prof. Pedro Alberto Barbetta, Dr.

_____

Prof. Emil Kupek, Dr.

_____

Prof. Jorge Coelho, Dr.

"Não se deve ir atrás de objetivos fáceis.
É preciso buscar o que só pode ser alcançado
por meio dos maiores esforços."
Albert Einstein.

Agradecimentos

Primeiramente eu tenho que agradecer aos meus pais, Silvia e Cláudio, os quais sempre me apoiaram, ajudaram, incentivaram e nunca me permitiram desistir perante as dificuldades que apareceram. Sem eles eu jamais teria concluído com sucesso esta dissertação. Aproveito para agradecer meus irmãos, Fábio, Fernando e Álvaro, que sempre confiaram em mim e acreditavam que este objetivo poderia ser alcançado.

Me considerado uma pessoa abençoada por ter duas mães chamadas Silvia e nascidas em Belém. A primeira é biológica, a segunda é acadêmica. A profª Silvia me ajudou inúmeras vezes antes, durante e depois do mestrado. Cresceu uma amizade e uma admiração tão forte, que eu sinto que ela representa mais do que a figura de uma orientadora, mais também de uma verdadeira mãe, em toda profundidade que esta palavra possui. Sem ela eu jamais teria obtido sucesso.

Agradeço meu orientador profº Paulo Freitas, pelas oportunidades nos projetos de pesquisa, pela orientação e pelos churrascos. Agradeço o profº Masanao e profº Barbetta pelas contribuições. Agradeço também ao Profº Fernando Cruz pela oportunidade concedida em ter minha primeira experiência na docência superior.

Aos meus amigos do laboratório LEA, em especial ao meu grande amigo Marcelo Tenório. O conheci logo na minha chegada a Florianópolis e me ajudou sempre que precisei. A Beatriz, nós ingressamos juntos no mestrado, e foram muitas as conversas, sobre disciplinas, congressos, projetos de pesquisa, enfim, uma grande amiga. Agradeço a Jaqueline, participamos de disciplinas e projetos juntos, uma amiga em todas as horas. Ao Gustavo, Rafael e o Vilson pelas conversas pela parceria no laboratório. Uma contribuição importante tiveram os amigos que fiz fora da universidade, em especial o Marcão e o Flávio, amizade que levarei pro resto da vida. O Limão e Rodrigo, amigos das baladas. Ao Lucas, amigo dentro e fora da UFSC.

Por último agradeço a minha maravilhosa namorada Gabriela, que surgiu na minha vida muito cedo, mas que apenas no final do mestrado formamos um casal. Ela me ajudou muito com incentivos e mais do que isso, quase todas as figuras desta dissertação foram feitas por ela, uma grande profissional. Além de toda essa ajuda, ainda me deu um filho lindo chamado Caio. Ambos sem dúvida me deram força para terminar esse mestrado. Amo vocês.

A todos que direta ou indiretamente me ajudaram a finalizar esta pesquisa.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Quadros

Lista de Siglas

Va	variável aleatória
Fdp	função de densidade de probabilidade
Fdc	função de distribuição acumulada
MMC	Método de Monte Carlo
GNA	Gerador de Números Aleatórios
FGVA	Função Geradora de Variáveis Aleatórias
TEC	Tempo Entre Chegadas
UML	Unified Modeling Language

Resumo

Com o intuito de melhorar o processo de descoberta de modelos teóricos de probabilidade nas situações em que a quantidade de dados para análise é insuficiente ou inexistente, é apresentado nesta pesquisa um processo que aplica o conhecimento tácito de especialistas de um domínio e o conhecimento teórico de estatísticos para extrair a informação sobre o comportamento de variáveis aleatórias contínuas. Os modelos teóricos considerados são: Uniforme, Triangular, Normal, Lognormal, Exponencial e Weibull. Neste processo alguns aspectos são os diferenciais dos métodos atuais, como a escolha da distribuição baseada nas suas características em relação à forma, a utilização de recursos visuais, o uso de uma linguagem familiar às pessoas sem muito conhecimento estatístico, aplicação para qualquer variável contínua, e para o cálculo dos parâmetros de cada distribuição, é requisitado ao especialista do domínio que forneça apenas dois valores do comportamento da variável. Os resultados da descoberta da forma da distribuição foram validados por especialistas do domínio, que interagiram com o software implementado com o processo utilizado. A validação do cálculo dos parâmetros foi realizada por estatísticos e por simulação.

Palavras chave: Reconhecimento de Padrões, Distribuição Contínuas de Probabilidade e Elicitação do Conhecimento Tácito.

Abstract

With the purpose of improve the process of theoretical probability discovery in situations in situations which the amount of data for analysis is insufficient or the data do not exist, it is presented in this research a process that uses tacit knowledge of experts in a domain and theoretical knowledge of statisticians to extract information about the behavior of continuous variables. The theoretical models are: Uniform, Triangular, Normal, Lognormal, Exponential and Weibull. In this process some aspects are the differential from the usual methods, like the distribution choice based on its shape characteristics, the use of visual components, the use of familiar language to people with little statistical knowledge, the application to any continuous variable and it is required from the expert in the domain only two values of the distribution behavior for the parameters calculation for each probability distribution. The results of the distribution shape discovery were validated by domain experts who used the software implemented with the described process. The parameter calculations were validated by statisticians and by simulation.

Keywords: Pattern Recognition, Continuous Probability Distribution and Knowledge Elicitation

Capítulo 1

Introdução

1.1 Motivação

A necessidade de informação é absolutamente crucial para todos os segmentos, desde para as grandes empresas até para assuntos pessoais. Dentre as várias formas presentes atualmente em conseguir se manter informado está a descoberta de padrões de comportamento.

Com esta informação é possível utilizar as técnicas de simulação. Estas técnicas, segundo FREITAS FILHO (2001), "permitem descrever o comportamento do sistema, construir teorias e hipóteses considerando as observações efetuadas e usar o modelo para prever o comportamento futuro, isto é, os efeitos produzidos por alterações no sistema ou nos métodos empregados em sua operação".

Esta informação também é de extrema serventia na construção de cenários dinâmicos para acompanhar e gerenciar atividades. Estes cenários são uma eficaz ferramenta pró-ativa na difícil tarefa de análise de riscos (FLORES et al., 2006).

Unificando a simulação de sistemas com análise de riscos é possível tornar mais eficiente a tarefa de tomadas de decisões, onde é importante também conhecer a natureza do evento, expressa em distribuição de probabilidade, para obter melhores resultados. Como benefício dessa união surge a possibilidade de descrever, entender e prever comportamentos de processos físicos complexos (GARTHWAITE et al., 2005).

Usualmente, estas informações estão contidas nas bases de dados. Estas bases registram o comportamento da atividade ou da variável, formando assim um histórico sobre o evento. A partir desses dados é possível descobrir o modelo teórico de probabilidade dessa variável.

Os modelos teóricos de probabilidade são funções matemáticas que representam probabilisticamente variáveis aleatórias que estão contidas no nosso cotidiano (TENÓRIO, 2005). Essas variáveis podem representar qualquer atividade, podendo ser qualitativa ou quantitativa. Apenas esta última está sob foco neste estudo.

O conhecimento desses modelos, que representam o comportamento de uma variável, é muito importante, pois a partir do momento que se pode representar esta variável com um modelo conhecido, é possível se valer dos benefícios oferecidos pela simulação de sistemas utilizando vários cenários.

Este é um exemplo da importância de se conhecer o modelo, existindo outras funcionalidades para tal informação. Dada essa necessidade, existem os testes de aderência, que servem justamente para o reconhecimento de modelos teóricos de probabilidade.

Estes testes podem ser aplicados para variáveis discretas e contínuas. Dentre estes testes podem-se destacar o Teste Qui-quadrado, Teste de Kolmogorov-Smirnov, Teste de Lilliefors, Teste de Anderson-Darling, Teste Tenório-Nassar (JANKAUSKAS & MCLAFFERTY, 1995, ROMEU, 2003, n.4, n.5 e n.6, NIST/SEMATECH, 2008, TENÓRIO, 2005).

Entretanto, para o sucesso de qualquer um desses testes, é indispensável a presença de histórico de dados, pois estes testes precisam verificar, por meio de comparações, se os dados seguem algum modelo teórico de probabilidade conhecido. Somente desta forma é possível fazer o reconhecimento de padrões pelos testes de aderência.

Logo, em casos onde a quantidade de dados para análise é insuficiente ou os dados simplesmente não existem, nenhum dos testes de aderência pode ser aplicado. Nessas situações é necessário encontrar alguma forma alternativa de extrair as informações.

Para suprir essa deficiência na descoberta dos modelos teóricos de probabilidade, existem processos capazes de extrair essas informações do comportamento da variável de interesse a partir do conhecimento tácito dos especialistas em um domínio.

Entende-se por conhecimento tácito, o conhecimento que o especialista possui em decorrência de sua experiência, mas que não consegue expressá-lo de um modo formal. A formalização do conhecimento do especialista, nesse estudo, resultará em um modelo teórico de probabilidade.

Entretanto, os processos atuais apresentam as seguintes características:

1. São, em sua maioria, direcionados a áreas específicas de atuação, onde já se tem a informação de qual modelo teórico de probabilidade melhor representa o comportamento da atividade, necessitando apenas dos parâmetros da distribuição;

2. Quando há a possibilidade de escolha da distribuição, cabe ao especialista o prévio conhecimento dos modelos teóricos de probabilidade;

3. Possuem uma linguagem estatística muito aprofundada de difícil entendimento, especialmente para pessoas que não possuem um contato maior com o campo da estatística;

4. Poucos apresentam um processo descrito detalhadamente por completo para aquisição das informações dos especialistas;

5. Não utilizam recursos visuais para auxiliar o processo de elicitação.

Com o intuito de aprimorar o procedimento da elicitação do conhecimento tácito de distribuições de probabilidade, é apresentado neste estudo um processo de elicitação baseado nas considerações da literatura atual referentes a boas práticas no processo de elicitação, juntamente com soluções inovadoras que atendam as preocupações levantadas anteriormente.

Desta forma, todos os benefícios providos com a descoberta do modelo poderão ser aplicados também aos casos em que não podem ser contemplados com os testes de aderência.

1.2 Objetivos

1.2.1 Objetivo Geral

O objetivo geral desta dissertação é desenvolver um processo capaz de extrair o conhecimento tácito de um especialista sobre o comportamento de uma atividade de seu domínio e formalizá-lo através de uma distribuição de probabilidade.

1.2.2 Objetivos Específicos

- Identificar e distinguir as características das distribuições contínuas de probabilidade utilizadas;
- Estabelecer um mecanismo de escolha da distribuição de acordo com as características das distribuições;
- Desenvolver cálculos capazes de obter os parâmetros das distribuições;
- Implementar o processo de elicitação para atender algumas variáveis aleatórias quantitativas contínuas;
- Utilizar recursos visuais para auxiliar o processo de elicitação;
- Validar o processo de elicitação proposto.

1.3 – Procedimentos Metodológicos

Esta pesquisa restringe sua aplicação às variáveis quantitativas contínuas, mais especificamente, às seguintes distribuições contínuas: Uniforme, Triangular, Normal, Lognormal, Exponencial e Weibull.

O resultado esperado do processo de elicitação da distribuição contínua de probabilidade deve possuir duas informações: a forma da distribuição e seus parâmetros.

Logo, o processo de elicitação deve ser composto de duas etapas. A primeira etapa é a descoberta da forma da distribuição de probabilidade e a segunda é o cálculo dos parâmetros desta distribuição.

As distribuições selecionadas foram analisadas e diferenciadas de acordo com suas características relacionadas à forma da distribuição. Estas características podem ser:

- De simetria em relação à média, podendo ser simétrica, assimétrica à esquerda ou assimétrica à direita;
- Do caráter assintótico da curva da distribuição.

Com a separação das distribuições de acordo com suas características, é possível formular perguntas de modo que as respostas possam apontar ao final do processo qual a distribuição de probabilidade mais indicada.

De posse da forma da distribuição de probabilidade, a próxima informação a ser obtida são os seus parâmetros. Algumas distribuições possuem parâmetros que não são intuitivos para pessoas que não tem um conhecimento específico em Estatística, como é o caso dos parâmetros α e β da distribuição Weibull.

Portanto, perguntas a respeito de valores que as pessoas de um modo geral tenham facilidade em responder são necessárias para que a partir delas, os parâmetros das distribuições possam ser calculados.

Outro ponto importante é manter o especialista do domínio focado durante todo o processo de elicitação. Para isto, é necessário utilizar mecanismos que chamem a sua atenção, desde o início do processo de elicitação.

Outro aspecto importante do processo de elicitação é manter este processo, que é de caráter fundamentalmente interativo, flexível ao usuário. Com isso, o usuário pode avançar ou retroceder nas etapas de elicitação, até atingir seu objetivo.

Artifícios visuais, como gráficos das distribuições, devem estar presentes no processo para facilitar o entendimento das informações fornecidas pelo Módulo de Elicitação.

Com o intuito de validar o processo de elicitação, um programa denominado de Módulo de Elicitação foi implementado computacionalmente. Esta validação foi divida em duas etapas. Para validar as questões das características das distribuições, especialistas nos mais variados domínios interagiram com o Módulo de Elicitação. Após esta etapa, os resultados obtidos foram analisados e as modificações sugeridas foram efetuadas no Módulo.

Na validação do cálculo dos parâmetros, a solução matemática foi apresentada para um grupo de estatísticos. Estes deram um parecer sobre os cálculos efetuados e sugeriram algumas alterações.

Adicionalmente à validação pelos estatísticos, um outro tipo de validação foi utilizado. Esta validação realizou o processo inverso ao cálculo dos parâmetros. A partir dos parâmetros calculados, foram encontrados os valores informados pelo especialista.

Caso os valores encontrados fossem iguais ou próximos aos valores que o especialista informou, isto significaria que o processo do cálculo dos parâmetros realmente estava adequado.

1.4 – Estrutura da Dissertação

Esta dissertação está divida em seis capítulos, sendo que o conteúdo de cada um será especificado a seguir.

O capítulo 1 faz uma introdução ao estudo realizado, situando o leitor no contexto do problema de pesquisa. Ainda aponta os objetivos geral e específicos, bem como cita os procedimentos, que serão detalhados no decorrer do trabalho, utilizados na elaboração da solução desenvolvida.

Após esta introdução, é realizada no capítulo 2 uma revisão de importantes conceitos estatísticos e computacionais utilizados neste estudo. O conhecimento destes conceitos é importante para a compreensão do texto.

De posse do conhecimento do problema e dos conceitos necessários, no capítulo 3 são especificados e detalhados os procedimentos que foram utilizados para elicitar modelos contínuos de probabilidade.

O capítulo 4 contém a documentação e as telas do software desenvolvido para aplicar a solução encontrada no capítulo anterior. Os fluxos de interação entre o software e o usuário são formalmente descritos em diagramas.

O penúltimo capítulo (capítulo 5) apresenta os testes e resultados da validação da solução encontrada nesta dissertação.

Por último, no capítulo 6 são discutidos os resultados encontrados no capítulo 5 e apresentada algumas considerações a respeito de todo o processo de elicitação, desde sua concepção, passando pelas dificuldades até se chegar aos resultados finais. Finalizando este último capítulo constam algumas sugestões de prosseguimento desta pesquisa, assinaladas como trabalhos futuros.

Capítulo 2

Revisão da literatura

Neste capítulo são apresentadas algumas definições importantes para o entendimento do estudo realizado. Deste modo, além de uma revisão destes conceitos, são apresentados também alguns trabalhos semelhantes que foram desenvolvidos nesta linha de pesquisa.

2.1 – Simulação de Sistemas Computacionais

A primeira definição que deve ser esclarecida ao falar sobre simulação é em relação ao modelo que se quer estudar. Portanto, um modelo é a descrição de algum sistema que se deseja prever o que acontecerá se certas ações forem tomadas (BRATLEY et al., 1987).

Quando esses sistemas são sistemas computacionais, a simulação consiste na "utilização de determinadas técnicas matemáticas, empregadas em computadores digitais, as quais permitem imitar o funcionamento de, praticamente, qualquer tipo de operação ou processo (sistemas) do mundo real" (FREITAS FILHO, 2001).

Existem várias aplicações interessantes onde à simulação de sistemas é utilizada, como por exemplo, casos em que: o sistema real ainda não existe, o experimento com o sistema real é dispendioso ou o experimento com o sistema real não é apropriado, como no caso de planejamento de atendimento de situações de emergência (FREITAS FILHO, 2001).

Para a construção desses modelos, é necessária a simulação da freqüência de ocorrência dos eventos envolvidos. No caso do atendimento de situações de emergência, é preciso estipular com que freqüência ocorre uma emergência e qual a duração do atendimento.

Como não existe um tempo exato entre as ocorrências das emergências e muito menos um tempo determinado destinado ao atendimento, o tempo destes acontecimentos, baseados na observação do modelo real, deve ser simulado no modelo computacional segundo uma variável aleatória.

A técnica para esta simulação é denominada de “Método de Monte Carlo”, e é apresentada na seção seguinte.

2.1.1 – Método de Monte Carlo

O Método de Monte Carlo (MMC) consiste em gerar dados artificialmente, empregando um gerador de números aleatórios (GNA) e uma distribuição de freqüência da variável de interesse (FREITAS FILHO, 2001).

O primeiro ponto do MMC se refere ao GNA, que segundo FREITAS FILHO (2001), “é um programa computacional que deve ser capaz de gerar valores aleatórios independentes uniformemente distribuídos em um intervalo de 0 a 1”.

Existem diversos algoritmos capazes de gerar tais números, mas estes números são chamados de pseudo-aleatórios. Isto decorre do fato da possibilidade de reprodução da seqüência de números produzidos por esse algoritmo.

Essa repetição da seqüência é possível porque a origem dos números é baseada no número anterior. Então o primeiro número é baseado em algum valor fornecido ao algoritmo, este valor inicial é usualmente chamado de “semente”.

Portanto, um mesmo algoritmo sendo iniciado com a mesma semente, produzirá sempre a mesma seqüência de números. Uma forma de conseguir gerar diferentes seqüências de números é utilizar como semente o número que o relógio do computador indica. Desta forma, baseado no horário da geração, uma seqüência de números gerada logo em seguida de outra produzirá seqüências diferentes.

Porém, nem todo número é satisfatório como “semente”, devido o fato de que após uma determinada quantidade de números (depende da qualidade da “semente”) a seqüência se repete. A escolha de uma boa “semente” é capaz de gerar uma seqüência incrivelmente grande de números pseudo-aleatórios independentes entre si.

De acordo com FREITAS FILHO (2001), uma análise estatística indica que a comparação de um conjunto de valores gerados artificialmente por um bom algoritmo e um conjunto de valores verdadeiramente aleatório, gerado pela natureza, não apresenta diferenças.

O segundo ponto importante no MMC está relacionado com a distribuição de freqüência da variável de interesse. Esta pode seguir um modelo teórico ou podem ser empíricas, não seguindo a nenhuma distribuição formalmente descrita estatisticamente.

Os valores, além de serem pseudo-aleatórios, podem ser gerados pelas funções geradoras de variáveis aleatórias (FGVA), baseados em uma dessas distribuições de probabilidade, desde que sejam fornecidos os parâmetros necessários. Estes parâmetros variam de acordo com a distribuição de probabilidade escolhida.

O funcionamento de um modelo utilizando o MMC ocorre da seguinte maneira: Primeiramente é carregada a função geradora de números aleatórios (GNA) juntamente com diversas FGVA. Cada distribuição de probabilidade possui sua própria FGVA. Portanto, na maioria dos modelos computacionais, para a utilização de uma determinada distribuição de probabilidade, basta indicar o nome da distribuição e fornecer os seus parâmetros característicos (FREITAS FILHO, 2001).

A Fig. 2.1 ilustra um exemplo de como funciona a interação entre o modelo computacional com as FGVA's. A sigla TEC significa o tempo entre chegadas, denotando o intervalo de tempo desde a última ocorrência, o surgimento de uma emergência, por exemplo, até a próxima ocorrência.

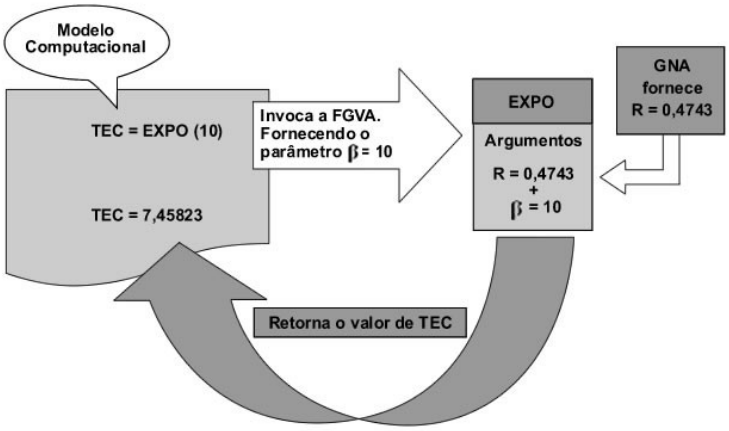

Figura 2.1 – Interação entre o modelo computacional e as FGVA's. Extraído de FREITAS FILHO (2001).

Neste modelo computacional, que reflete um modelo real, o tempo entre a chegada de um novo evento foi modelado como sendo uma distribuição Exponencial com média igual a 10.

Assim, FGVA da Exponencial é acionada juntamente com seu parâmetro. A FGVA da Exponencial recebe o número aleatório do GNA. Com base neste número aleatório e no valor de seu parâmetro, o valor da geração é encontrado e é retornado para a variável TEC no modelo computacional.

2.2 – Variáveis Aleatórias

O autor BARBETTA (2006), chama de **variáveis "**as características que podem ser observadas (ou medidas) em cada elemento da população, sob as mesmas condições".

Geralmente em qualquer tipo de experimento, os resultados deste são relacionados com um número, sendo estabelecida uma regra de associação entre eles. Esta regra de associação é denominada variável aleatória (va). É variável porque pode assumir diversos valores numéricos, e aleatória devido o fato do valor observado depender do resultado do experimento sob incerteza por aleatoriedade (DEVORE, 2006).

Como exemplo desta associação, pode-se citar o experimento em que uma pessoa tenta acessar uma área restrita de segurança com um cartão magnético. Quando esta pessoa passa o cartão magnético na leitora de cartão, esta verifica se a pessoa é autorizada (S) ou não é autorizada (N) a ingressar naquela área.

Portanto, sendo δ o espaço amostral, conforme ilustrado na Fig. 2.2, os possíveis resultados são: $\delta = \{S, N\}$. Uma va é definida como: X(S) = 1 e X(N) = 0. A va X indica se a pessoa pode (1) ou não (0) acessar a área restrita.

Para DEVORE (2006), "Em termos matemáticos, uma va é uma função cujo domínio é o espaço amostral e o contra-domínio é um conjunto de números reais".

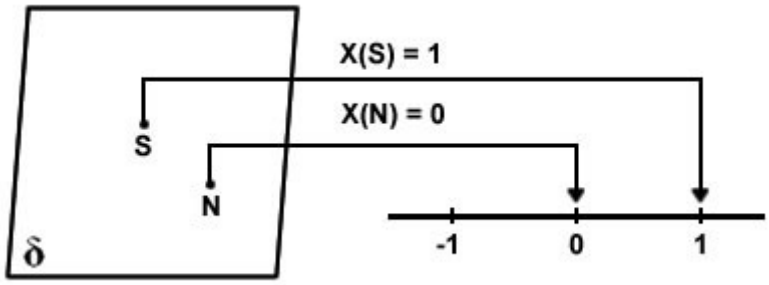

Figura 2.2 – Mapeamento para os números reais.

Neste exemplo, cada elemento da va X que consta no espaço amostral foi relacionado explicitamente com um número, sendo os únicos valores possíveis 0 e 1. Porém, existir apenas dois valores não é uma regra, podem existir tantos valores quantos forem necessários. Mas no caso de existirem apenas o 0 e 1, tais variáveis receberam

uma nomenclatura especial devido ao primeiro indivíduo que as estudou, são chamadas de variáveis aleatórias de Bernoulli (DEVORE, 2006).

Nos casos em que os resultados possíveis de uma variável são números de uma certa escala, é dito que esta variável é quantitativa. Quando os resultados possíveis são atributos ou qualidades, a variável é considerada qualitativa BARBETTA (2006).

No contexto das variáveis quantitativas, estas podem ser divididas em contínuas ou discretas. A Fig. 2.3 apresenta graficamente esta divisão.

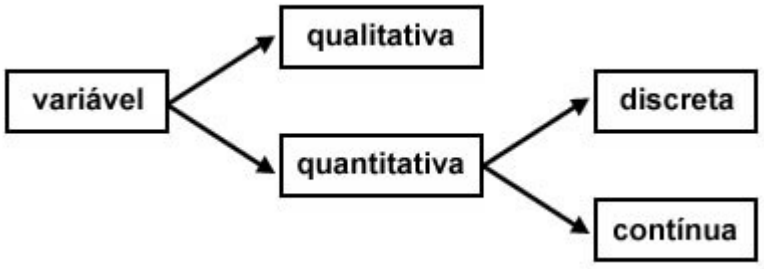

Figura 2.3 – Classificação da variável em termos do nível de mensuração.

Este estudo se restringe apenas às variáveis aleatórias contínuas, por este motivo não será tecida uma abordagem mais profunda das variáveis aleatórias discretas.

2.2.1 – Variáveis Aleatórias Discretas

“São as variáveis que só assumem valores que podem ser listados” (BARBETTA, 2006).

O autor DEVORE (2006) escreveu a seguinte definição, “uma variável aleatória discreta é uma variável, cujos valores possíveis constituem um conjunto finito ou podem ser relacionados em uma seqüência infinita na qual haja um primeiro elemento, um segundo elemento e assim por diante”.

O exemplo da área de segurança mencionado anteriormente é de uma variável aleatória discreta. Pois o espaço amostral $\delta = \{0,1\}$ constitui um conjunto finito.

Um outro exemplo de variável aleatória discreta, que se encaixe na segunda parte da definição, é um experimento onde se deseja saber o número total de rodadas de lançamento de uma moeda para que três indivíduos obtenham resultados iguais. Cada rodada é formada pelo lançamento de três moedas, uma para cada indivíduo.

Este experimento não tem um conjunto finito, no entanto o espaço amostral é $\delta = \{1,2,3,...\}$, ou seja, até que ocorra este resultado, haverá a primeira rodada com tentativa dos três indivíduos obterem os mesmos resultados, a segunda rodada e assim por diante. As variáveis discretas geralmente resultam de alguma contagem (BARBETTA, 2006).

2.2.2 – Variáveis Aleatórias Contínuas

Diferentemente das variáveis aleatórias discretas, as variáveis aleatórias contínuas costumam ser obtidas por uma mensuração e podem assumir qualquer valor num intervalo (BARBETTA, 2006).

Pela definição encontrada em DEVORE (2006), "uma variável aleatória é dita contínua se o seu conjunto de valores possíveis consistir do intervalo completo de todos os valores, isto é, para cada A < B, qualquer valor x entre A e B for possível".

Por esta definição é possível identificar inúmeros exemplos de variáveis aleatórias contínuas, para citar um exemplo, pode-se pensar no tempo necessário para perfuração de um poço de petróleo. Ainda de acordo com DEVORE (2006), se a escala de medida puder ser subdividida em quantas partes forem desejadas, esta variável é contínua. Portanto este é caso da variável tempo de perfuração, pois este tempo pode ser subdividido em anos, meses, dias, horas, minutos, segundos e assim por diante.

2.2.2.1 – Função Densidade de Probabilidade

A distribuição de probabilidades de uma variável aleatória contínua *X* pode ser descrita por uma função densidade de probabilidade (fdp) (MONTEGOMERY & RUNGER, 2003). De acordo com BARBETTA et al. (2004), para que *f*(x) seja uma fdp legítima, as seguintes condições devem ser respeitadas:

6. $f(x) \geq 0, \forall x \in \Re$ e
7. $\int_{-\infty}^{\infty} f(x)dx = 1$

Se $A = [a,b]$, então $P(A) = \int_{a}^{b} f(x)dx$

Isto significa que a probabilidade de *X* ter um determinado valor no intervalo [a,b] é a área contida entre este intervalo e situada abaixo da curva da função densidade *f*(x). Esta curva é denominada de curva de densidade (MOORE & MCCABE, 2002). A Fig. 2.4 ilustra um exemplo da curva de densidade.

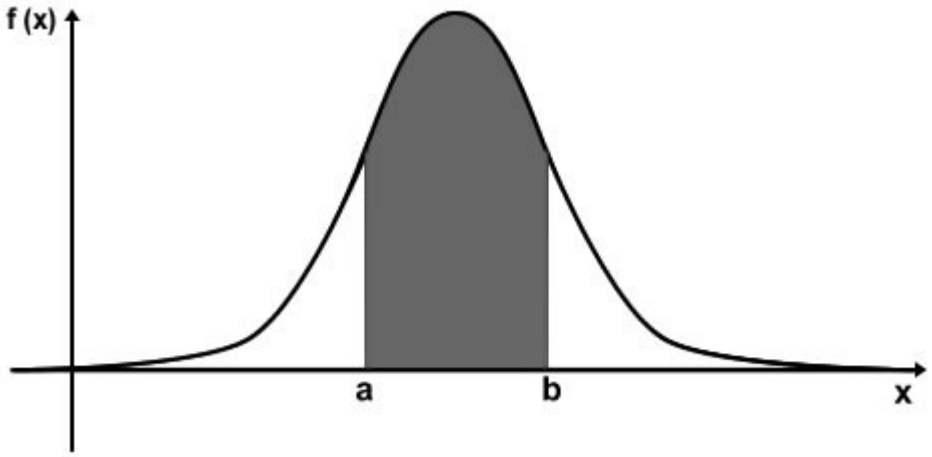

Figura 2.4 – P($a \leq X \leq b$).

"Se *X* é uma variável aleatória contínua , então para qualquer valor *c*, P($X = c$) = 0. Além disso, para quaisquer dois valores *a* e *b* com $a < b$, P($a \leq X \leq b$) = P($a < X \leq b$) = P($a \leq X < b$) = P($a < X < b$)" (DEVORE, 2006).

Isso significa que a probabilidade de qualquer valor específico é zero e a probabilidade de um intervalo independe da inclusão de seus valores extremos.

2.2.2.2 – Função de Distribuição Acumulada

A função de distribuição acumulada (fda) é um método alternativo para descrever uma variável aleatória (MONTEGOMERY & RUNGER, 2003).

A fda de uma va *X* fornece, para qualquer valor de x, a probabilidade P($X \leq$ x),

isto é: $F(x) = \int_{-\infty}^{x} f(x)dx$, $\forall x \in \Re$ (BARBETTA et al., 2004).

Para cada x', F(x') é a área abaixo da curva de densidade e à esquerda de x', conforme ilustrado na Fig. 2.5.

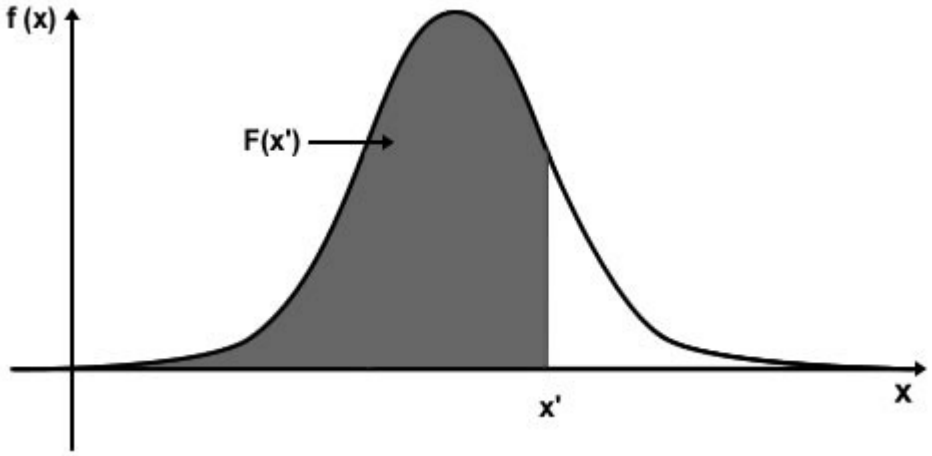

Figura 2.5 – fdp: P ($X \leq$ x') = $\int_{-\infty}^{x'} f(x)dx$.

A Fig. 2.6 apresenta a fda equivalente a fdp da Fig. 2.5. Em uma análise das duas figuras se pode perceber que com um x' contendo grande parte da área sob a curva de densidade, a probabilidade de um valor estar neste intervalo é próxima de 1.

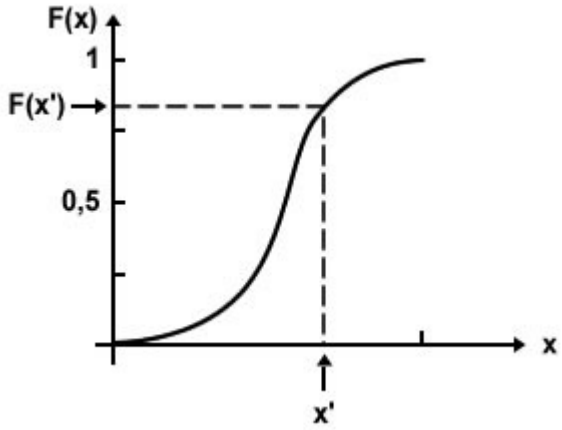

Figura 2.6 – fda: P ($X \leq$ x') = F(x').

Segundo BARBETTA et al. (2004), é possível obter qualquer probabilidade por meio da função de distribuição acumulada.

Então, para qualquer número a e b com $a < b$:

- P($X < a$) = P($X \leq a$) = F(a)
- P($X > b$) = 1 – F(b)
- P($a \leq X \leq b$) = F(b) – F(a)

Desta forma, obtendo-se a fda, a tarefa de calcular qualquer probabilidade envolvendo *X* pode ser simplificada, não necessitando o uso de integração (DEVORE, 2006).

Dada a fda de uma va contínua, a fdp pode ser determinada por:

$$f(x) = \frac{d}{dx} F(x)$$

para todo ponto *x* em que *F* é derivável (MONTEGOMERY & RUNGER, 2003, BARBETTA et al., 2004).

2.2.2.3 – Percentis

Os percentis compreendem a divisão de um conjunto de dados em cem partes iguais, sendo úteis para especificar uma medida de posição (LARSON & FARBER, 2004).

O conceito de percentil pode ser melhor compreendido quando comparado com o uso de porcentagem. No seguinte exemplo é dito que 80% dos estudantes ficaram com nota inferior a nota do estudante A. Isto significa que a pontuação na prova deste estudante A estava no percentil 80 ou 80º percentil da população de estudantes.

Na definição de DEVORE (2006), seja *p* um número entre 0 e 1. O **100*p*-ésimo percentil** da distribuição de uma variável aleatória contínua *X*, representada por $\eta(p)$ é definido por:

$$p = F(\eta(p)) = \int_{-\infty}^{\eta(p)} f(x)dx$$

Conforme a expressão na definição, o mesmo autor complementa, "$\eta(p)$ é o valor no eixo das medidas tal que 100*p*% da área abaixo do gráfico de *f*(x) está à esquerda de $\eta(p)$ e 100(1 – *p*)% está à direita". Desta forma η (0,80), o 80º percentil, é tal que a área abaixo do gráfico de *f*(x) à esquerda de η (0,80) é 0,80.

Um importante conceito neste contexto de percentis diz respeito ao 50º percentil ou percentil 50 ou mediana. A mediana de uma distribuição contínua, ou seja η (0,50),

significa que metade da área abaixo da curva de densidade está à esquerda e metade à direita deste valor.

Se o gráfico da fdp à esquerda de um ponto é espelho do gráfico à direita desse ponto, então se diz que é simétrica a fdp desta variável aleatória contínua. Isto significa que ela possui a mediana igual ao ponto de simetria, já que metade da área abaixo da curva encontra-se em cada lado desse ponto.

2.3 – Distribuições de Probabilidade

As distribuições contínuas de probabilidade consideradas neste estudo são: Uniforme, Triangular, Normal, Lognormal, Exponencial e Weibull. Estas distribuições foram escolhidas por possuírem características que representam um grande número de variáveis. Todas são detalhadas nas próximas seções.

2.3.1 – Distribuição Uniforme

A distribuição Uniforme possui a característica de apresentar a mesma probabilidade de ocorrência para todos os valores compreendidos dentro de um intervalo, com um valor *mínimo* e um *máximo*.

Ela é tradicionalmente utilizada quando a única informação que se dispõe é o valor mínimo e o máximo que uma variável aleatória pode apresentar. Isso indica um desconhecimento do fenômeno aleatório sob análise (FREITAS FILHO, 2001).

Ainda de acordo com o mesmo autor, a sua grande importância está em gerar valores aleatórios entre 0 e 1. Esta tarefa é fundamental para a geração de variáveis aleatórias pelos diversos métodos numéricos existentes.

Como visto anteriormente, a área de simulação de sistemas se apóia bastante na geração de valores aleatórios uniformemente distribuídos (BRATLEY et al., 1987).

DEFINIÇÃO: Diz-se que uma va X é uniformemente distribuída em $a \leq x \leq b$ se sua função densidade de probabilidade é (SPIEGEL, 1978, LAW & KELTON,1991):

$$f(x;a,b) = \begin{cases} \dfrac{1}{b-a} & \text{se } a \leq x \leq b \\ 0 & \text{caso contrário} \end{cases}$$

A função de distribuição acumulada é dada por:

$$F(x;a,b) = \begin{cases} 0 & \text{se } x < a \\ \dfrac{x-a}{b-a} & \text{se } a \leq x \leq b \\ 1 & \text{se } b < x \end{cases}$$

Fonte: (SPIEGEL, 1978).

No Quadro 2.1 são apresentadas quatro expressões referentes às medidas estatísticas: *média, mediana*, *moda* e *variância*.

Quadro 2.1 – Características da distribuição Uniforme

Média	$\dfrac{a+b}{2}$
Mediana	$\dfrac{a+b}{2}$
Moda	Não existe *moda.*
Variância	$\dfrac{(b-a)^2}{12}$

Fonte: (LAW & KELTON, 1991).

A Fig. 2.7 apresenta um exemplo do modelo de uma distribuição Uniforme com o valor *mínimo* igual a 1 e o valor *máximo* igual a 10.

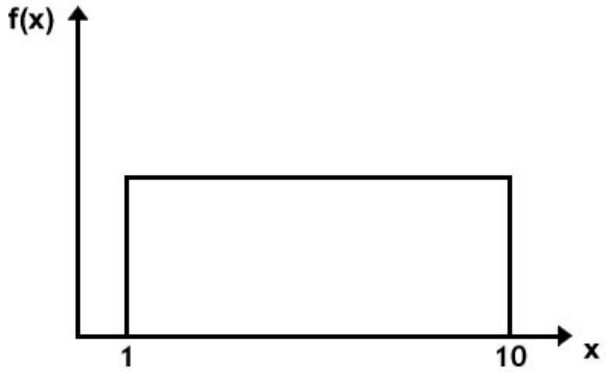

Figura 2.7 – Distribuição Uniforme, *mínimo* = 1 e *máximo* = 10.

2.3.2 – Distribuição Triangular

A exemplo da distribuição Uniforme, apresentada na seção anterior, a distribuição Triangular é utilizada principalmente quando não se sabe muito a respeito da curva associada à variável aleatória. Um outro parâmetro, além dos limites superior (b) e inferior (a), deve ser informado. Este terceiro parâmetro se refere ao valor mais provável (c) (FREITAS FILHO, 2001).

A distribuição Triangular oferece perspectivas de melhor aderência para modelos reais do que aqueles com base na distribuição Uniforme. (FREITAS FILHO, 2001).

DEFINIÇÃO: Diz-se que uma variável aleatória *X* possui uma distribuição Triangular com parâmetros $a < c < b$, se a função densidade de probabilidade é:

$$f(x; a, c, b) = \begin{cases} \dfrac{2(x-a)}{(b-a)(c-a)} & \text{se } a \leq x \leq c \\ \dfrac{2(b-x)}{(b-a)(b-c)} & \text{se } c < x \leq b \\ 0 & \text{caso contrário} \end{cases}$$

A função de distribuição acumulada é dada por:

$$F(\mathrm{x};\, a,\, c,\, b) = \begin{cases} 0 & \text{se } \mathrm{x} < a \\ \dfrac{(x - a)^2}{(b - a)(c - a)} & \text{se } a \leq \mathrm{x} \leq c \\ 1 - \dfrac{(b - x)^2}{(b - a)(b - c)} & \text{se } c < \mathrm{x} \leq b \\ 1 & \text{se } \mathrm{x} > b \end{cases}$$

Fonte: (LAW & KELTON, 1991).

No Quadro 2.2 são apresentadas quatro expressões referentes às medidas estatísticas: *média, mediana, moda* e *variância*.

Quadro 2.2 – Características da distribuição Triangular

Média	$\dfrac{a + b + c}{3}$
Mediana	$a + \dfrac{\sqrt{(b - a)(c - a)}}{\sqrt{2}}$ se $c \geq \dfrac{b - a}{2}$ $b - \dfrac{\sqrt{(b - a)(b - c)}}{\sqrt{2}}$ se $c \leq \dfrac{b - a}{2}$
Moda	c
Variância	$\dfrac{a^2 + b^2 + c^2 - ab - ac - bc}{18}$

Fonte: (SPIEGEL, 1978).

A Fig. 2.8 ilustra a forma do modelo de uma distribuição Triangular, com *mínimo* igual a 0*, mod*a igual a 5 e *máximo* tendo o valor 10.

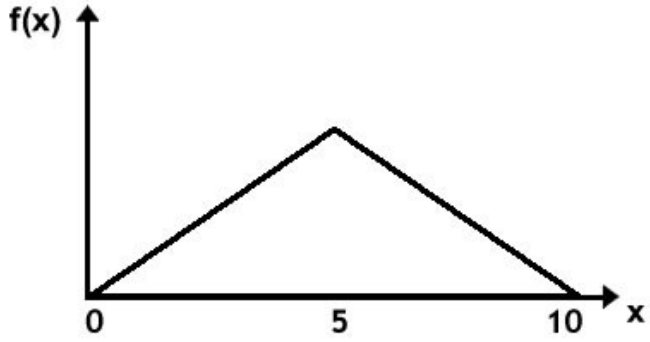

Figura 2.8 – Distribuição Triangular com *mínimo* = 0, *moda* = 5 e *máximo* = 10.

2.3.3 – Distribuição Normal

A distribuição Normal é tida como a distribuição mais importante de todas em Probabilidade e Estatística pela alta aplicabilidade (DEVORE, 2006). Esta distribuição tem como parâmetros a *média* (μ) e o *desvio padrão* (σ). A *média* indica a posição central da distribuição e o *desvio padrão* refere-se à dispersão da distribuição. Nesta distribuição, "o valor de σ é a distância de μ até os pontos de inflexão da curva (os pontos em que a curva muda de direção)" (DEVORE, 2006).

A curva Normal tem forma de sino e o valor da *média* coincide com os valores da *mediana* e da *moda,* já que a curva Normal é simétrica em torno da *média*. Estes valores estão localizados bem no centro da curva, que é o lugar em que a curva é mais alta (MOORE & MCCABE, 2002).

Por causa desta simetria pode-se dizer que a distribuição de valores menores que a *média* e a distribuição de valores maiores que a *média* são perfeitamente proporcionais, ou seja, o lado da distribuição dos valores abaixo da *média* é a imagem especular do lado da distribuição acima da *média*.

Outra característica importante desta distribuição esta relacionada às extremidades da curva, que neste caso é assintótica, ou seja, as extremidades da curva se estendem indefinidamente pelo eixo x em direção ao infinito (positivo ou negativo), sem jamais tocar no eixo (LARSON & FARBER, 2004).

A maior probabilidade de ocorrência está localizada em torno da *média* e à medida que os valores se distanciam desta *média*, a probabilidade vai decrescendo (MONTEGOMERY & RUNGER, 2003). A Fig 2.9 mostra os valores percentuais de ocorrência que correspondem a cada intervalo. Estes intervalos são formados de acordo com a quantidade de *desvio padrão* que os valores estão distantes da *média*.

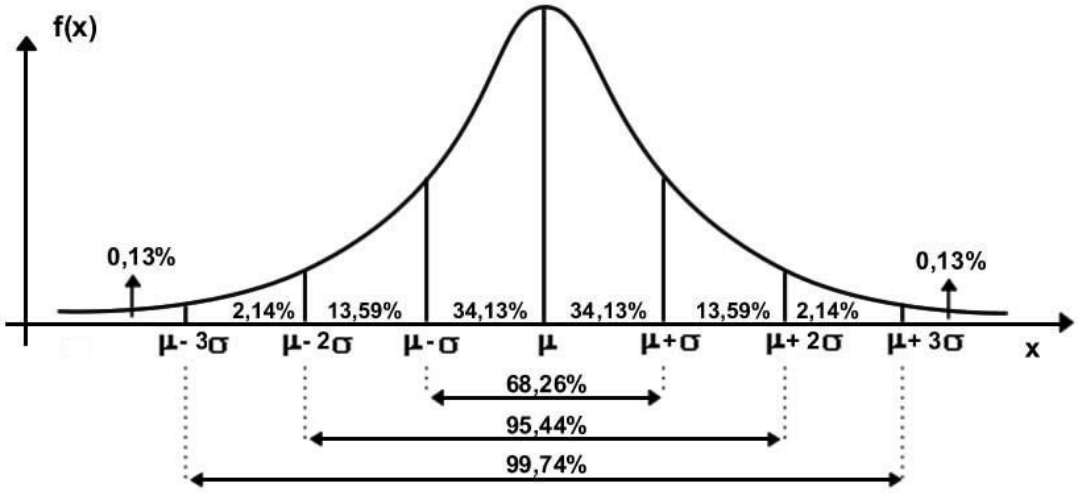

Figura 2.9 – Concentração de valores em torno da *média*.

A definição encontrada em DEVORE (2006) atesta que:

DEFINIÇÃO: Uma variável aleatória contínua *X* possui uma distribuição Normal com parâmetros μ e σ, onde $-\infty < \mu < \infty$ e $\sigma > 0$, se a fdp de *X* for:

$$f(x;\mu,\sigma) = \begin{cases} \dfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} & -\infty < x < \infty \end{cases}$$

A função de distribuição acumulada correspondente é:

$$F(x) = P(X \leq x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

Fonte: (SPIEGEL, 1978)

No Quadro 2.3 são apresentadas as representações das medidas estatísticas: *média, mediana, moda* e *variância*. Como foi mencionado anteriormente, os valores de *média, mediana* e *moda* são exatamente os mesmos.

Quadro 2.3 – Característica da distribuição Normal

Média	μ
Mediana	μ
Moda	μ
Variância	σ^2

Fonte: (SPIEGEL, 1978)

De acordo com os valores de *média* e *desvio padrão* a distribuição Normal sofre alterações em sua forma. Estas mudanças, no entanto, não mudam a característica básica de simetria da distribuição.

Com o aumento do *desvio padrão* o comportamento esperado da distribuição é que haja um achatamento da curva, com uma maior dispersão em torno da *média*, como é observada na curva (c) da Fig. 2.10.

Se a *média* permanecer a mesma e o *desvio padrão* diminuir, apresentará um maior valor de freqüência e uma acentuada concentração de valores será observada bem próximo à *média*. O que pode ser observado nas curvas (b) e mais acentuado ainda na curva (a).

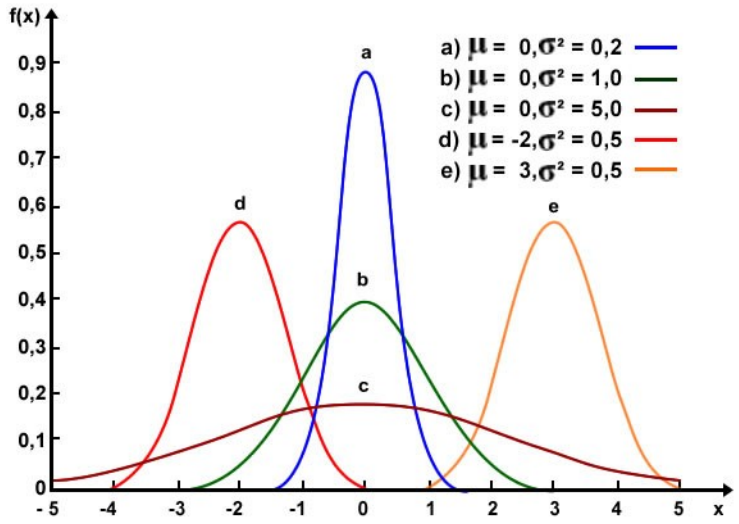

Figura 2.10 – Distribuições normais com μ e σ diferentes.

Já distribuições que possuem o mesmo *desvio padrão*, mas com valores de *média* diferentes, terão a mesma dispersão, ou seja, o mesmo formato da curva. O que irá variar será apenas a localização da média no eixo x. Haverá um deslocamento da curva de acordo com o valor da *média*, como pode ser observado nas curvas (d) e (e).

Para o cálculo da probabilidade P(a $\leq X \leq$ b) quando X é uma variável aleatória com distribuição Normal, com parâmetros μ e σ, deve-se ser efetuado o seguinte cálculo:

$$P(a \leq X \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \qquad 2.1$$

Porém, nenhuma das técnicas de integração padrão pode ser usada para calcular a expressão 2.1 (DEVORE, 2006). Para contornar esta situação, para os parâmetros μ = 0 e σ = 1 a expressão 2.1 foi calculada e tabulada para valores determinados de "a" e "b". Esta tabela permite também o cálculo da probabilidade de qualquer outro valor de μ e σ, desde que seja pensado como quantidade de σ que X está distante de μ.

Quando a distribuição Normal possui os parâmetros μ = 0 e σ = 1, a distribuição é chamada de distribuição Normal Padrão (MOORE & MCCABE, 2002, BARBETTA et al., 2004, MONTEGOMERY & RUNGER, 2003, LARSON & FARBER, 2004).

"A distribuição Normal Padrão freqüentemente não serve como modelo para uma população natural. Ao contrário, é uma distribuição de referência a partir da qual podem ser obtidas informações sobre outras distribuições normais" (DEVORE, 2006).

Assim, qualquer problema relacionado a uma distribuição Normal pode ser pensado em termos da distribuição Normal Padrão.

O valor equivalente a variável aleatória *X,* que segue uma distribuição Normal, na distribuição Normal Padrão é dado por:

$$Z = \frac{X - \mu}{\sigma} \qquad 2.2$$

Com a expressão 2.2, o valor calculado de Z é a distância que o valor de X está da média em termos de σ.

O cálculo da probabilidade P(a $\leq X \leq$ b) que segue uma distribuição Normal qualquer, é a área sob a curva desses dois valores, dado que "as probabilidades de uma

variável com distribuição Normal podem ser representadas por áreas sob a curva da distribuição Normal Padrão" (BARBETTA et al., 2004).

2.3.4 – Distribuição Lognormal

Neste item apresenta-se a definição encontrada em DEVORE (2006) desta distribuição e algumas expressões de suas medidas estatísticas.

DEFINIÇÃO: Diz-se que uma variável aleatória não-negativa *X* possui uma distribuição Lognormal se a variável aleatória $Y = \ln(X)$ possui uma distribuição Normal. A fdp resultante de uma variável aleatória Lognormal quando $\ln(X)$ tiver distribuição Normal com parâmetros μ e σ é:

$$f(x;\mu,\sigma) = \begin{cases} \dfrac{1}{\sigma x\sqrt{2\pi}} e^{-\frac{[\ln(x)-\mu]^2}{2\sigma^2}} & \text{se } x > 0 \\ 0 & \text{caso contrário} \end{cases}$$

Conforme o autor LAW & KELTON (1991), esta distribuição não possui expressão fechada para a função de distribuição acumulada.

No Quadro 2.4 são apresentadas quatro expressões referentes às medidas estatísticas: *média, mediana, moda* e *variância.*

Quadro 2.4 – Característica da distribuição Lognormal

Média	$e^{\left(\mu + \frac{\sigma^2}{2}\right)}$
Mediana	e^{μ}
Moda	$e^{\left(\mu - \sigma^2\right)}$
Variância	$e^{\left(2\mu + \sigma^2\right)}\left(e^{\sigma^2} - 1\right)$

Fonte: (LAW & KELTON, 1991).

A distribuição Lognormal possui uma relação direta com a distribuição Normal, como foi apresentado anteriormente na definição desta distribuição. Portanto qualquer distribuição Lognormal tem sua respectiva forma na distribuição Normal.

Assim como a distribuição Normal, a *variância* ou *desvio padrão* desta distribuição diz respeito ao parâmetro de forma. Enquanto que o parâmetro *média* corresponde a escala.

Pode-se observar na Fig. 2.11 que com o valor do *desvio padrão* pequeno, há uma simetria em torno do valor da *moda* (curva d). À medida que o *desvio padrão* aumenta, fica mais evidente uma assimetria à direita (curva a).

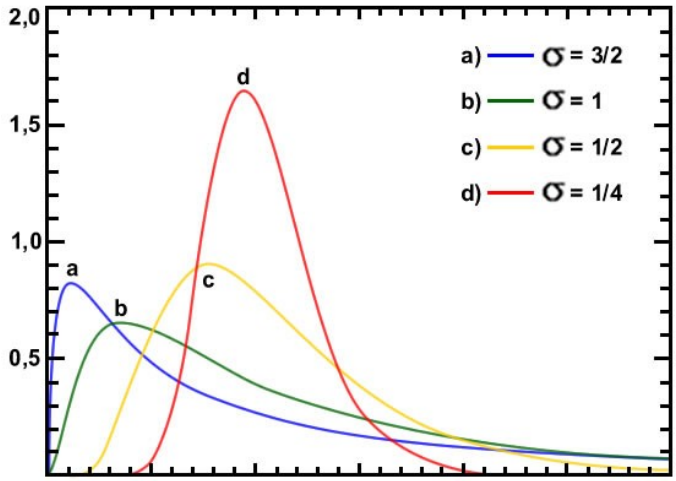

Figura 2.11 – Curvas da distribuição Lognormal com vários valores de σ .

A distribuição Lognormal não permite valores negativos, logo seu limite inferior é no mínimo igual a zero. Já a outra extremidade da curva é assintótica.

2.3.5 – Distribuição Exponencial

A distribuição Exponencial possui este nome devido à presença da função Exponencial na função de densidade de probabilidade (MONTEGOMERY & RUNGER, 2003). A seguir é apresentada a definição encontrada em DEVORE (2006).

DEFINIÇÃO: Diz-se que X tem uma distribuição Exponencial com parâmetro λ ($\lambda > 0$) se a fdp de X for:

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{se } x \geq 0 \\ 0 & \text{caso contrário} \end{cases}$$

A função de distribuição acumulada é:

$$F(x;\lambda) = \begin{cases} 1 - e^{-\lambda x} & \text{se } x \geq 0 \\ 0 & \text{caso contrário} \end{cases}$$

Fonte: (DEVORE,2006).

No Quadro 2.5 são apresentadas quatro expressões referentes às medidas estatísticas: *média, mediana, moda* e *variância.*

Quadro 2.5 – Característica da distribuição Exponencial

Média	$\frac{1}{\lambda}$
Mediana	$\frac{\ln 2}{\lambda}$
Moda	0
Variância	$\frac{1}{\lambda^2}$

Fonte: (MONTGOMERY, 2003).

A distribuição Exponencial possui uma propriedade que nem uma outra distribuição contínua de probabilidade apresenta, a propriedade da falta de memória. O conhecimento relativo aos resultados ocorridos anteriormente não afeta as probabilidades de eventos futuros (MONTGOMERY, 2003).

Esta distribuição pressupõe a independência entre os eventos, ou seja, $P(X < t_1 + t_2 \mid X > t_1) = P(X < t_2)$.

Portanto esta distribuição não é adequada para casos onde a informação de eventos passados influencia na probabilidade de eventos futuros, como, por exemplo, modelar o tempo até uma falha de um equipamento, dado que este equipamento sofre desgaste com o uso.

A distribuição Exponencial é bastante utilizada na modelagem de tempos decorridos entre dois eventos, especialmente quando estes são causados por um grande número de fatores independentes (FREITAS FILHO, 2001, BARBETTA et al., 2004).

De acordo com o valor do parâmetro λ, algumas mudanças são notadas na curva desta distribuição.

Pode ser observado na Fig. 2.12 que a diminuição de λ acarreta em uma maior dispersão dos valores. Enquanto que valores maiores de λ concentram a probabilidade de ocorrência em valores mais baixos.

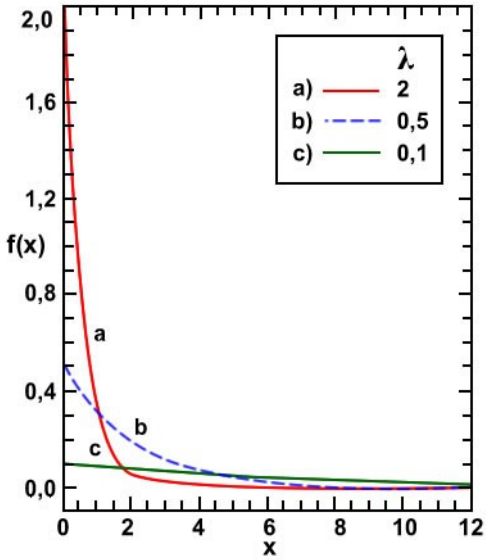

Figura 2.12 – Distribuição Exponencial com vários valores de λ. Extraído de MONTGOMERY (2003).

A distribuição Exponencial pode seguir para uma distribuição Weibull quando $X^{\frac{1}{\alpha}}$. A distribuição Weibull será apresentada na seção seguinte.

2.3.6 – Distribuição Weibull

A família de distribuições Weibull possui dois parâmetros, que são α e β, onde α diz respeito à forma e β à escala da distribuição. Porém existem situações práticas aonde o valor mínimo da variável pode não ser zero, mas outro valor chamado de δ. A quantidade δ pode ser vista então como um terceiro parâmetro da distribuição, o que

foi feito por Weibull em seu trabalho original. Deste modo, se δ = 5, todas as curvas da distribuição são deslocadas cinco unidades para a direita (DEVORE, 2006).

DEFINIÇÃO: Uma variável aleatória *X* tem distribuição Weibull com parâmetro de forma α e parâmetro de escala β, para α >0 e β >0, se a fdp de *X* for:

$$f(x;\alpha,\beta)=\begin{cases}\dfrac{\alpha}{\beta^{\alpha}}x^{\alpha-1}e^{-\left(\frac{x}{\beta}\right)^{\alpha}} & \text{se } x \geq 0\\ 0 & \text{caso contrário}\end{cases}$$

A função de distribuição acumulada é:

$$F(x;\alpha,\beta)=\begin{cases}1-e^{-\left(\frac{x}{\beta}\right)^{\alpha}} & \text{se } x \geq 0\\ 0 & \text{caso contrário}\end{cases}$$

Fonte: (DEVORE, 2006).

No Quadro 2.6 são apresentadas quatro expressões referentes às medidas estatísticas: *média, mediana, moda* e *variância*.

Quadro 2.6 – Característica da distribuição Weibull

Média	$\dfrac{\beta}{\alpha}\Gamma\left(\dfrac{1}{\alpha}\right)$
Mediana	$\beta\ \ln(2)^{\frac{1}{\alpha}}$
Moda	$\begin{cases}\beta\left(\dfrac{\alpha-1}{\alpha}\right)^{\frac{1}{\alpha}} & \text{se } \alpha>1\\ 0 & \text{caso contrário}\end{cases}$
Variância	$\dfrac{\beta^{2}}{\alpha}\left\{2\Gamma\left(\dfrac{2}{\alpha}\right)-\dfrac{1}{\alpha}\left[\Gamma\left(\dfrac{1}{\alpha}\right)\right]^{2}\right\}$

Fonte: (LAW & KELTON, 1991).

Nas expressões da *média* e *variância* existe a presença da função *Gama* (Γ). Esta função possui a seguinte definição:

DEFINIÇÃO: A função Gama $\Gamma(\alpha)$ é definida por $\int_0^\infty x^{\alpha-1}e^{-x}dx$, para $\alpha > 0$ (DEVORE, 2006).

A distribuição Weibull possui a característica de assumir diversos formatos de distribuições diferentes de acordo com o valor de α, que é um parâmetro de forma da distribuição (DEVORE, 2006). O parâmetro β, que se trata de um parâmetro de escala, dependendo do valor ele comprime ou expande a distribuição no eixo horizontal, referente aos valores. A seguir é apresentado o comportamento da distribuição Weibull em relação à variação de α.

Se $\alpha = 1$, a fdp é reduzida à distribuição Exponencial. De forma que a distribuição Exponencial torna-se um caso especial das distribuições Weibull (DEVORE, 2006), como exemplificado na Fig. 2.13.

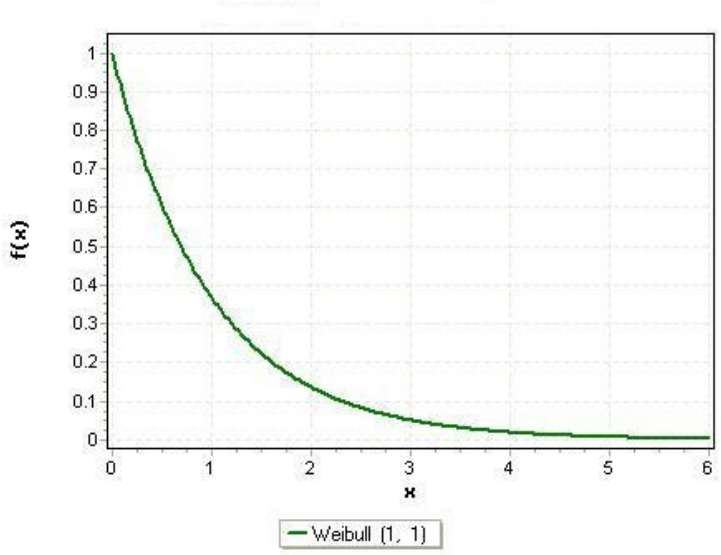

Figura 2.13 - Distribuição Weibull com $\alpha = 1$.

Com $\alpha = 2$, a distribuição Weibull se assemelha a distribuição Lognormal, com uma assimetria a direita, como pode ser visto na Fig. 2.14.

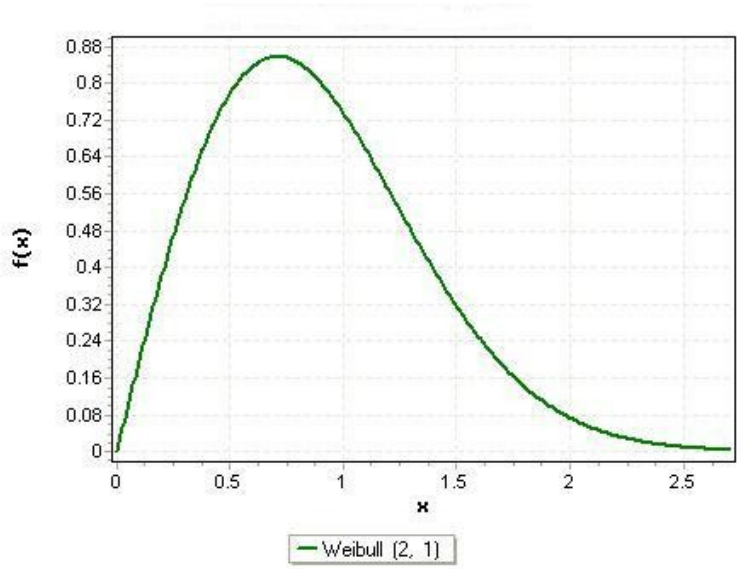

Figura 2.14 – Distribuição Weibull com α = 2.

No caso de α = 3, a moda da distribuição Weibull se posiciona mais próxima da média, porém com alguma assimetria a direita ainda. Com esse parâmetro a forma começa a ficar simétrica, aproximando-se da distribuição Normal, o que é observado na Fig. 2.15.

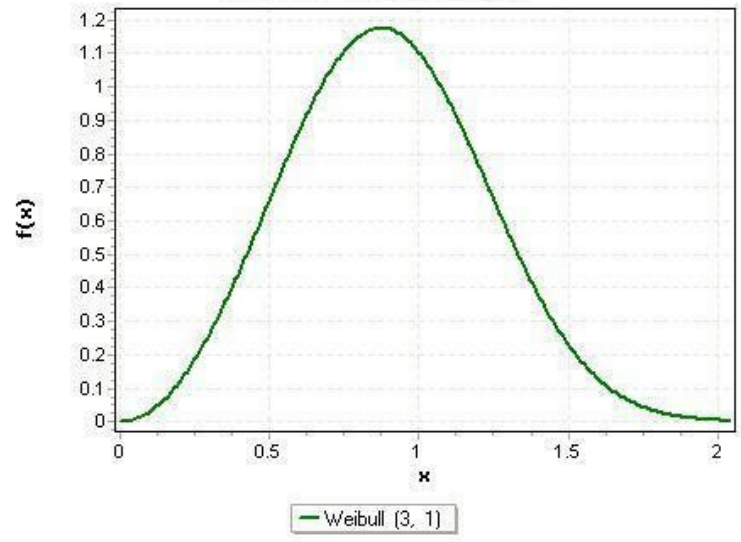

Figura 2.15 – Distribuição Weibull com α = 3.

A partir de α = 4 (Fig. 2.16), a distribuição Weibull, que apesar do valor da moda ainda estar perto da média, começa a ter a forma que é adotada como forma característica desta distribuição nesta pesquisa, assimétrica à esquerda.

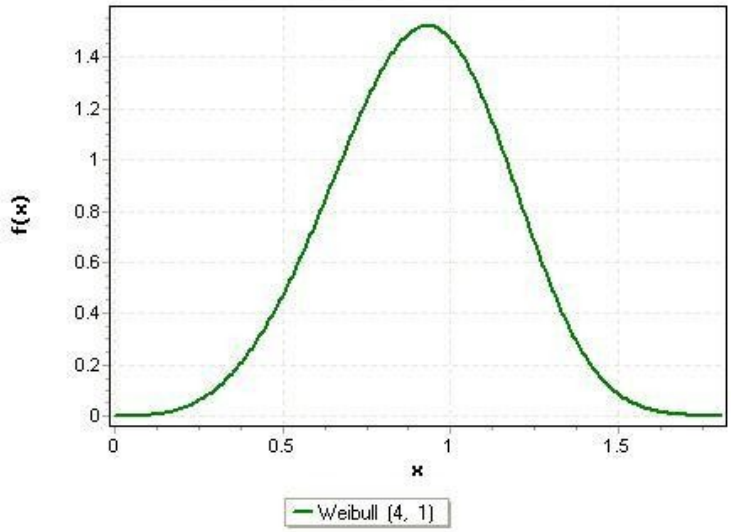

Figura 2.16 – Distribuição Weibull com $\alpha = 4$.

Na Fig. 2.17, que possui $\alpha = 5$, comprova-se que aumentando o valor de α a forma da distribuição fica cada vez mais assimétrica à esquerda. Desta forma, com o valor de $\alpha = 4$ passa a ser referência no processo de descoberta dos parâmetros desta distribuição, tendo em vista que existe a obrigatoriedade deste parâmetro ser superior a 4.

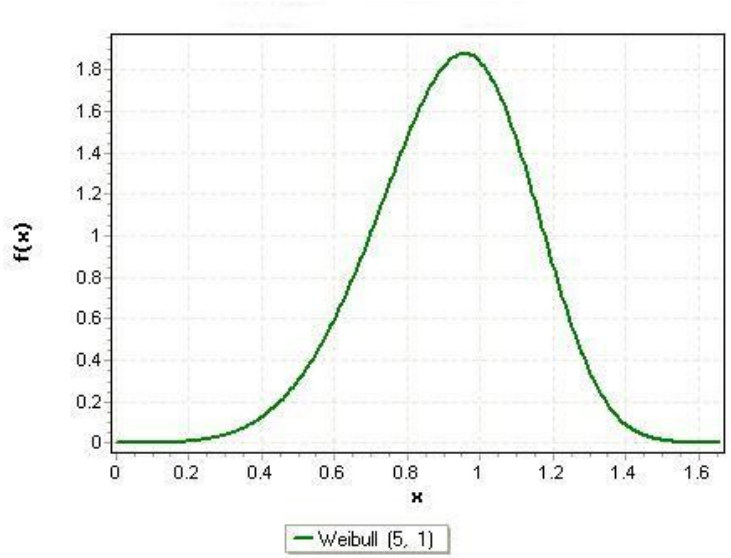

Figura 2.17 – Distribuição Weibull com $\alpha = 5$.

2.4 – Elicitação do Conhecimento

O termo elicitação vem do inglês *elicitation*, o que significa, segundo o dicionário eletrônico BABYLON (2008), extrair, obter, retirar, produzir. Outro dicionário eletrônico, o THE FREE DICTIONARY (2008) amplia a definição anterior, se referindo ao termo elicitação como a extração de alguma coisa latente ou chegar em uma verdade, por exemplo, por meio da lógica.

Aplicando este conceito para o estudo dos modelos de distribuições de probabilidade realizado nesta dissertação, o autor GARTHWAITE et al. (2005) apresenta sua própria definição. Segundo ele, elicitação é o processo de formular o conhecimento e crença de uma pessoa sobre uma ou mais quantidades incertas em uma distribuição de probabilidade.

Para JENKINSON (2005), sabendo-se que o conceito de probabilidade é algo subjetivo, pessoal de cada indivíduo que considera um evento, elicitação constitui-se da tarefa de converter as idéias da mente de um indivíduo em um número real entre zero e um.

Como apresentado anteriormente, elicitação é o processo de formular em termos probabilísticos as crenças de uma pessoa. Mas geralmente o termo, pessoa, é denominado de especialista.

O uso do termo especialista sugere que esta pessoa possua um alto grau de conhecimento sobre um determinado domínio que as demais pessoas, em sua grande maioria, não possuem.

Existem também, segundo GARTHWAITE et al. (2005), outros usos de elicitação, onde o especialista possui pouca ou nenhuma especialidade no significado comum dessa palavra. O autor cita o caso de tomadas de decisão, em relação a comportamento de risco, de adolescentes. Poderia ser perguntado a um adolescente, como ele vê esses riscos. Neste caso, o ponto central do estudo é justamente a falta de conhecimento do "especialista". Portanto, a forma mais simples de descrever um especialista é como a pessoa a qual se deseja elicitar o conhecimento.

É conveniente ter uma pessoa que ajude na tarefa de formular o conhecimento do especialista na forma probabilística. Esta pessoa é denominada de "facilitadora". No caso em que o especialista é um estatístico, ou tiver uma familiaridade grande com os termos estatísticos, a presença desse facilitador não parece ser necessária, o que na prática é algo muito raro.

Além do exposto acima, o mesmo autor insiste que a tarefa de elicitação é um processo complexo que exige habilidades para ser bem feito, e um facilitador tem um importante papel para a obtenção desse resultado.

Para a obtenção de uma elicitação bem sucedida, é necessário ter em mente a clara distinção entre a qualidade do conhecimento do especialista e a precisão com que o conhecimento é traduzido para uma forma probabilística. Portanto, uma elicitação é bem realizada quando a distribuição encontrada é o mais próximo possível do conhecimento do especialista, sem levar em consideração a qualidade deste conhecimento (GARTHWAITE et al., 2005).

O mesmo autor dividiu o processo de elicitação em quatro estágios separados, são eles:

- *Configurar: Este estágio consiste na preparação para a elicitação, selecionando e treinando os especialistas, identificando quais aspectos do problema se deve elicitar.*

- *Elicitar: É o principal estágio do processo. Extrai as informações necessárias para a descoberta da distribuição, em relação aos aspectos observados no estágio anterior.*

- *Ajustar: Neste estágio as informações colhidas no estágio 2 são analisadas de forma que possam representar alguma distribuição de probabilidade.*

- *Adequado: A elicitação é, invariavelmente, um processo iterativo. Portanto, este último estágio se refere a adequação da distribuição encontrada com o conhecimento do especialista. Caso haja a necessidade, deve-se retornar ao estágio 2 e recomeçar o processo de elicitação.*

A Fig. 2.18 representa graficamente o processo descrito.

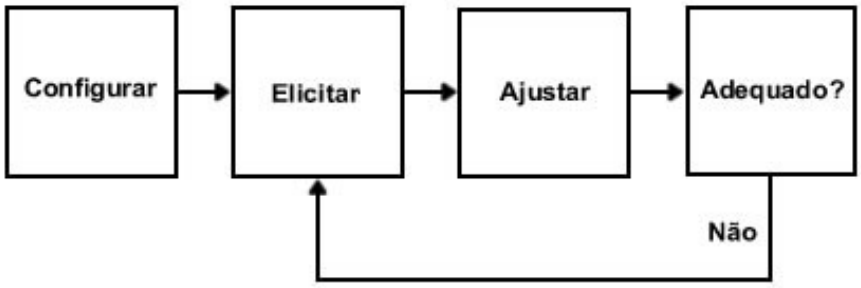

Figura 2.18 – Processo de elicitação. Extraído de (GARTHWAITE et al, 2005).

A primeira ação a ser executada em um processo de elicitação, segundo LOVERIDGE (2004), é a utilização de questionários ou formulários apropriados para a aquisição de informação. Neste estudo as perguntas do questionário serão aplicadas pelo software módulo de elicitação.

A formulação desses questionários pode ser dada de uma forma em que nem sempre é necessário que todas as questões sejam respondidas. Existe um encadeamento nas questões, e dependendo da situação, apenas algumas perguntas se fazem necessárias. Este é o caso da estruturação em forma de árvores de decisão, o qual também é aplicado nesta dissertação.

Serão apresentados, a seguir, alguns aspectos relevantes que devem ser observados no momento da formulação desses questionários.

2.4.1 – Formulação do questionário

A formulação de um questionário é uma etapa longa que deve ser conduzida com o maior cuidado, sendo necessário definir claramente os objetivos e a população a que se aplica. Por estas razões BARBETTA (2006) chamou a atenção para alguns procedimentos na formulação de um questionário:

- *Separar as características a serem levantadas.*
- *Fazer uma revisão bibliográfica para verificar como mensurar adequadamente algumas características.*

- *Estabelecer a forma de mensuração das características (variáveis) a serem levantadas.*
- *Elaborar uma ou mais perguntas para cada característica a ser observada.*
- *Verificar se a pergunta está suficientemente clara.*
- *Verificar se a forma da pergunta não está induzindo alguma resposta.*
- *Verificar se a resposta não é óbvia.*

A seguir serão expostas algumas considerações na formulação de questionários abordadas por AYYUB (2001). A construção de um questionário deve começar por definir sua relevância nos seguintes níveis:

1. *Relevância do estudo para os participantes (pessoas que podem ser afetadas ou podem afetar o processo de elicitação): É muito importante deixar claro para estas pessoas, o objetivo do processo de elicitação e qual a importância e relevância deste processo para elas. Com isso, objetiva-se elevar tanto o grau de atenção, como o nível de sinceridade.*

2. *Relevância das questões para o estudo: Cada questão ou tópico no questionário precisa dar suporte para o objetivo do estudo. A clareza desta relevância é fundamental para reforçar a confiabilidade dos dados coletados dos participantes.*

3. *Relevância das questões para os participantes: Cada questão ou tópico do questionário precisa ser claramente relevante para cada participante, especialmente quando o questionário é aplicado para participantes com visões diferentes e conhecimentos diferentes.*

A seguir são apresentadas orientações na construção das questões:

- *Cada item do questionário deve incluir uma questão apenas. É uma péssima prática incluir duas questões em uma.*
- *Questões ou declarações não devem ser ambíguas. O uso de termos ambíguos também deve ser evitado.*

- *O nível do vocabulário deve ser o menos técnico possível. Questões longas devem ser evitadas. As palavras devem ser escolhidas cuidadosamente para atingir o objetivo do estudo da maneira mais confiável.*
- *O uso de questões factuais é preferível, em detrimento de questões mais abstratas. Questões que são referentes a questões concretas e específicas resultam em respostas concretas e específicas.*
- *Questões devem ser formuladas cuidadosamente para evitar vícios dos participantes. As questões devem ser apresentadas de forma neutra, sem induzir o participante a uma determinada resposta.*

As questões podem ser classificadas em dois tipos: abertas e fechadas. As questões fechadas possuem as seguintes características:

- *Limita a possibilidade de respostas.*
- *Pode fornecer orientação ao participante, facilitando o processo.*
- *Oferece respostas completas.*
- *Permite lidar com questões mais sensíveis e delicadas.*
- *Permite a comparação das respostas dos participantes.*
- *Produz respostas que podem ser facilmente codificadas e analisadas.*
- *Podem ser capciosas.*
- *Permite que participantes ignorantes escolham uma resposta por adivinhação.*
- *Pode frustrar o participante com a apresentação de somente respostas inapropriadas em sua opinião.*
- *Limita a escolha das respostas possíveis.*
- *Não permite a detecção de variação nas interpretações das questões pelos participantes.*
- *Resulta em uma pequena variação artificial, devido a limitada possibilidade de respostas.*

As questões abertas possuem as características a seguir:

- *Não limita a possibilidade de repostas.*
- *É adequada para questões sem respostas conhecidas.*
- *É adequada para questões que possuam muitas respostas possíveis.*

- *É preferível para tratar com assuntos complexos.*
- *Permite a expressão da criatividade.*
- *Pode resultar em informações irrelevantes.*
- *As respostas dos participantes podem não ser padronizada, dificultando a comparação das respostas.*
- *Pode produzir dados que são difíceis de serem codificados e analisados.*
- *Requer uma habilidade de escrita significativa.*
- *Pode não expressar apropriadamente as dimensões e complexidade do assunto.*
- *Requer mais tempo dos participantes.*
- *Podem ser vistas como difíceis de responder, ocasionando no desestímulo por parte do participante para responder todas corretamente.*

O número de questões e a ordem de apresentação também são um fator importante na elaboração de um questionário. Alguns cuidados devem ser tomados, são eles:

- *Questões sensíveis ou abertas devem ser abordadas no final do questionário.*
- *O questionário deve começar com questões simples, que são mais fáceis de responder.*
- *A ordem lógica das questões deve ser formulada de modo que as questões do início do questionário proporcionem embasamento para a resposta das questões do final do questionário.*
- *As questões devem seguir outras ordens lógicas, tais como a ordem cronológica ou o fluxo do processo.*
- *Questões com tipos e formatos diferentes devem ser misturadas, para manter o interesse do participante.*
- *A ordem das questões pode estabelecer um funil, começando com perguntas mais gerais e especializando aos poucos. Entretanto, esta técnica pode não ser apropriada para todos os casos. É necessário um cuidado especial para analisar para cada caso se esta técnica pode ser aplicada.*

O estágio final da preparação de um questionário é a formulação de um documento introdutório que contenha o objetivo do estudo e estabeleça a sua relevância.

A seguir, são apresentadas algumas das dificuldades e armadilhas na elaboração de um questionário segundo AYYUB (2001). São sugeridas pelo autor algumas soluções.

1. *Os participantes podem achar que o questionário não é legítimo e que possui intenções desconhecidas escondidas no texto. Um bom documento introdutório com a explicação de tudo que se espera do questionário pode ajudar a afastar esse sentimento.*

2. *Os participantes podem achar que o resultado do questionário pode ser usado contra eles. Tópicos delicados que não sejam extremamente necessários e questões duplicadas devem ser removidos. Algumas vezes assegurar o anonimato é necessário.*

3. *Os participantes podem não querer responder ao questionário com medo de revelar sua ignorância no assunto. Enfatizar que não existem respostas certas ou erradas e assegurar o anonimato, pode ser necessário.*

4. *O participante pode fornecer respostas padronizadas, ou seja, respostas que ele ache que estão sendo desejadas no questionário. Assuntos delicados desnecessários e questões duplicadas devem ser removidos. Talvez seja necessário também assegurar o anonimato.*

5. *O participante pode pensar que o questionário é uma perda de tempo. Treinamento e educação podem ajudar a dar, ao participante, um comportamento diferente em relação ao questionário.*

6. *Os participantes podem achar alguma questão muito vaga e que não pode ser respondida. Esta questão deve então ser reformulada para que fique mais clara.*

O questionário deve ser completo, abrangendo todas as características necessárias para alcançar o seu objetivo e não conter nenhuma pergunta que não esteja relacionada com esse propósito. Esta coesão é importante pelo fato de que quanto mais

longo for o questionário, menor tende a ser a qualidade e confiabilidade das respostas (BARBETTA, 2006).

Antes de iniciar a aplicação do questionário ao público alvo, é necessário verificar se o mesmo está corretamente formulado. Para isto, BARBETTA (2006) sugere que a realização de um *pré-teste* é fundamental.

Este *pré-teste* consiste em aplicar o questionário para alguns indivíduos que possuam características semelhantes aos indivíduos da população em estudo.

É somente nesta etapa que é possível detectar possíveis falhas que tenham passadas desapercebidas no processo de formulação das perguntas. Além das dificuldades e armadilhas expostas por AYYUB (2001), BARBETTA (2006) destaca:

- *Ambigüidade de alguma pergunta.*
- *Resposta que não havia sido prevista.*
- *Não variabilidade de respostas em alguma pergunta, entre outras.*

A aplicação do questionário não necessita ser obrigatoriamente por meio de um formulário de papel. Na pesquisa desenvolvida neste estudo, o questionário foi aplicado na forma de uma interface de um software, o qual foi desenvolvido especialmente com esta finalidade.

Outro aspecto fundamental é o planejamento de como usar as respostas das várias perguntas para se chegar a algum resultado conclusivo da pesquisa (BARBETTA, 2006).

Para a estruturação lógica do encadeamento das perguntas desta pesquisa, foi utilizada uma estrutura denominada de árvores de decisão. Esta estrutura é descrita na seção seguinte.

2.4.2 – Árvores de Decisão

As árvores de decisão são freqüentemente utilizadas para examinar a informação disponível com a finalidade de se tomar uma decisão (AYYUB, 2001).

O funcionamento de uma árvore de decisão é basicamente receber como entrada um objeto ou situação descritos por um conjunto de atributos e retornar uma decisão. A saída apontada dependerá do valor de entrada (RUSSEL, 2004). Portanto, o ponto

final (decisão) da árvore de decisão é alcançado após a execução de uma seqüência de verificações.

Cada nó equivale a uma verificação, onde dependendo do resultado, outro nó pode ser verificado ou uma decisão é indicada. Existem pelo menos dois tipos de saída do nó. São do tipo booleano, em que apenas duas opções de resposta são possíveis, ou com múltiplas alternativas, havendo a possibilidade do encaminhamento para uma quantidade maior de nós.

Vale salientar que sempre apenas um caminho pode ser percorrido por vez, ou seja, a partir de um valor de entrada, o nó tomará a decisão de qual caminho único seguir. Cada caminho completo que começa na raiz, que é o primeiro nó de decisão, e segue até a folha, que é a última instância da árvore ou decisão a ser tomada, é caracterizado como um cenário possível de decisões com sua possível conseqüência (AYYUB, 2001).

2.5 – Trabalhos Correlatos

Nesta seção são apresentados algumas abordagens encontradas na literatura atual sobre elicitação de probabilidade.

No campo da descoberta de distribuições de probabilidades, existem alguns trabalhos que se baseiam em algum modelo teórico para a descoberta da distribuição e outros que não se valem de nenhuma indicação prévia de qual distribuição aquela determinada variável possa seguir.

O autor JENKINSON (2005) apresentou quatro métodos divididos em duas categorias que elucidam alguns métodos de descobrir a distribuição de probabilidade de uma variável sem considerar os modelos teóricos conhecidos.

A primeira categoria é conhecida como tendo métodos não paramétricos, nos quais o especialista não aproveita os conhecimentos dos modelos teóricos existentes no processo de descoberta da curva da variável. Qualquer curva pode ser encontrada nessa abordagem.

Esta categoria é composta pelo método denominado Amostra Prévia Equivalente - EPS (*Equivalent Prior Sample*) e o método Amostra Futura Hipotética – HFS (*Hypothetical Future Sample*). Ambos os métodos citados são utilizados para elicitar variáveis aleatórias que restringem seus resultados a faixa compreendida entre 0 e 1.

O escopo desta dissertação não se encaixa na categoria apresentada e sim na segunda categoria, que é descrita a seguir. Portanto, não será explicado o funcionamento do processo dos dois métodos citados.

A outra categoria é referente ao método paramétrico, no qual as informações obtidas do especialista serão ajustadas à curva de um modelo teórico a partir do conhecimento de uma dada distribuição contínua de probabilidade formalmente já definida. Os dois métodos representativos desta categoria são o método da função da distribuição acumulada (fda) e o segundo método da função de densidade de probabilidade (fdp).

No método fda é fornecido ao especialista várias probabilidades e depois pra cada probabilidade é requisitado que ele indique qual o valor da variável que não ultrapasse a probabilidade dada, ou seja, é necessário que o especialista seja capaz de atribuir vários percentis.

Dependendo da qualificação do especialista, essa abordagem se faz muito arriscada, dado que indicar um valor de percentil não é uma tarefa muito fácil. Indicar vários valores de percentis se torna a cada percentil mais arriscada, ainda mais que os especialistas geralmente possuem muito pouco conhecimento probabilístico.

O segundo método desta segunda categoria, o fdp, se fundamenta em representar graficamente a função de densidade. Sendo necessário que o especialista forneça inicialmente a média e a moda. Em seguida outros valores de percentis devem ser indicados, aperfeiçoando a aderência dos valores fornecidos com alguma curva de um modelo teórico.

Os quartis geralmente são perguntados ao especialista, desejando-se assim que o especialista divida a extensão em quatro partes iguais. Para isto, o método bastante utilizado é o método da bisecção.

Neste método, o especialista divide o intervalo de interesse pela metade. Após esta divisão é pedido ao especialista que adote este valor como sendo a mediana. Em seguida é separado um dos intervalos para análise. Sobre este intervalo é requisitado a mesma operação, que seja divida pela metade. Este valor se torna o quartil superior, caso o intervalo separado após a primeira divisão tenha sido o superior. Se o intervalo analisado foi o inferior, este valor é referenciado como o quartil inferior.

Esta operação pode ser refeita quantas vezes for necessária até que se tenha um nível de detalhamento adequado, obtendo-se vários valores de quantis.

Para a obtenção dos quantis, LOVERIDGE (2004) sugere que o especialista defina dois valores (um inferior e um superior) indicando que se ocorressem esses valores, o especialista ficaria “surpreso”. Estes valores seriam equivalentes aos quantis 10% e 90%.

Em seguida, seria requisitado a definição de outros dois valores (um inferior e um superior) indicando que se ocorressem esses valores, o especialista ficaria “perplexo”. Estes valores seriam considerados os quantis 99% e 1%.

Outra vez, assim como no primeiro método desta categoria, várias informações específicas da área da probabilidade são necessárias para o ajuste das informações do especialista para a curva de uma distribuição de probabilidade.

O objetivo desta dissertação será utilizar um método que utilize os conhecimentos de alguns dos modelos teóricos e poucas informações do especialista para encontrar a distribuição que melhor modela o comportamento da variável de interesse.

Além de reduzir a quantidade de informação necessária para a descoberta da distribuição de probabilidade, têm-se o intuito de simplificar o máximo possível as perguntas, trazendo o linguajar para perto de pessoas com pouco conhecimento estatístico e probabilístico.

Dada a importância da elicitação das distribuições de probabilidade, há uma grande variedade de aplicações específicas nas mais diversas áreas. O trabalho escrito por JENKINSON (2005) apresenta diversos exemplos.

Capítulo 3

Processo utilizado para a elicitação do conhecimento

Este capítulo aborda os detalhes do processo utilizado para a elicitação do conhecimento tácito dos modelos teóricos de distribuições contínuas de probabilidade.

O processo de elicitação utilizado é baseado nos procedimentos descritos por GARTHWAITE et al. (2005). Algumas modificações foram realizadas, considerando que estas deixariam o processo mais claro.

A Fig. 3.1 contém as etapas que devem ser percorridas até que se obtenha a informação final para apresentar ao especialista.

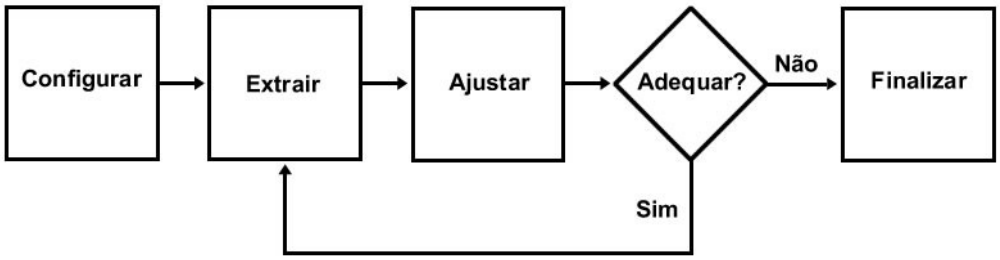

Figura 3.1 – Processo de elicitação utilizado.

A primeira etapa é a configuração, como foi descrito anteriormente, esta etapa consiste na preparação para elicitação, identificando quais aspectos do problema serão elicitados.

No processo desta pesquisa, esta etapa é composta primeiramente por uma explanação dos procedimentos e objetivos do processo de elicitação.

Em seguida aparece a identificação do especialista e da variável que se deseja elicitar. Esta identificação da variável tem como principal objetivo focar o especialista a aplicar os recursos do processo com aquela variável apenas, diminuindo a probabilidade de que durante a elicitação o especialista comece a pensar em outra variável. Dado que se ele começar o processo pensando em uma variável e terminar pensando em outra, a elicitação estará comprometida, tanto para a primeira quanto para a segunda variável.

É ainda nesta etapa que o especialista tem a opção de escolher se a variável que ele está pensando possui alguma característica dentre as encontradas na literatura. Esta

associação é feita devido à constante representação dessas características por uma mesma distribuição de probabilidade. Este procedimento será explicado com mais detalhes na seção 3.1, que trata da descoberta da forma da distribuição.

O último procedimento desta etapa são algumas questões que verificam se a variável escolhida é compatível com as limitações da solução desenvolvida. Salientando que o escopo deste trabalho se restringe às variáveis aleatórias contínuas.

Depois de devidamente configurado, as próximas etapas são: extração, ajuste e adequação. Estas três etapas não são separadas, elas são pequenas partes de um todo responsável pela aquisição da forma e parâmetros da distribuição.

Portanto, esta grande etapa contém dois ciclos. O primeiro é referente à descoberta da forma da distribuição contínua de probabilidade que melhor representa a atividade analisada pelo especialista. Na seção 3.1, é apresentada a metodologia aplicada para obter este objetivo.

Em complemento ao primeiro ciclo, após o conhecimento da forma da distribuição de probabilidade, é necessário calcular os parâmetros característicos desta distribuição. A seção 3.2 trata deste tópico, utilizando uma subseção para cada distribuição abordada nesse estudo, totalizando seis distribuições contínuas de probabilidade. Há a necessidade de tratar cada distribuição separadamente pela impossibilidade de existir uma solução comum a todas, dado que possuem diferentes características de forma e parâmetros.

Finalmente, depois de conhecido o modelo teórico de probabilidade, surge a última etapa do processo de elicitação, que finaliza o processo com um arquivo contendo todas as informações relevantes que foram extraídas no processo. Estas informações são: Nome do Especialista, domínio de aplicação, nome e descrição da variável, a característica encontrada na literatura (opcional), a distribuição de probabilidade e seus parâmetros.

Nas duas seções seguintes são descritos os detalhes dos procedimentos utilizados para a descoberta da forma da distribuição e do cálculo dos parâmetros.

3.1 – Procedimento para a descoberta da forma

A primeira etapa no processo de elicitação é a descoberta da forma da distribuição de probabilidade que melhor representa uma determinada variável.

Para alcançar este propósito, seis distribuições contínuas de probabilidade foram analisadas e diferenciadas de acordo com suas características. A seguir são apresentadas as particularidades de cada distribuição que são relevantes para a escolha da distribuição de probabilidade mais adequada. São elas:

- Uniforme: A distribuição Uniforme, como apresentado na seção 2.2, possui a característica de apresentar a mesma probabilidade de ocorrência para qualquer valor situado dentre um limite mínimo e máximo.

- Triangular: A distribuição Triangular possui a característica de simetria, apresentando a maior concentração dos seus valores em torno da média. Adicionalmente a isso, existem valores limites (mínimo e máximo) que eliminam a possibilidade de ocorrência de valores fora destes limites.

- Normal: Assim como a distribuição Triangular, a distribuição Normal também apresenta a maior concentração dos seus valores em torno da média. Entretanto, não existem valores limites. A probabilidade de ocorrência vai diminuindo indefinidamente à medida que os valores se afastam da média e valores muito pequenos ou muito grandes, apesar de não aparecerem freqüentemente, podem ocorrer.

- Lognormal: Esta distribuição tem a propriedade de assimetria à direita da média. A maior concentração dos valores está à esquerda da média. Apesar dos valores abaixo da média serem mais freqüentes, os menores valores possíveis não são os mais encontrados. As ocorrências começam com valores baixos, aumentam até atingirem um pico e após isto à medida que os valores vão se distanciando da média, a probabilidade de ocorrência decresce indefinidamente. Portanto, apesar de existir um valor mínimo, valores muito altos podem aparecer mesmo que esporadicamente.

- Exponencial: Da mesma forma que a distribuição Lognormal, a distribuição Exponencial também apresenta assimetria à direita da média. A diferenciação entre elas está na característica de ocorrência dos menores

valores. Na distribuição Exponencial, os menores valores possíveis são os mais freqüentes. À medida que os valores vão aumentando, a probabilidade de ocorrência vai diminuindo. Não existe um limite máximo, valores muito altos podem aparecer ocasionalmente.

- Weibull: A distribuição Weibull tem a característica de assimetria à esquerda da média. A maior concentração dos valores está à direita da média (nos maiores valores). As ocorrências dos valores vão aumentando à medida que valores mais altos aparecem até atingir um pico. Depois a probabilidade de ocorrência começa a decrescer indefinidamente, tendo valores muito altos como ocorrências isoladas.

- Beta: Diversas formas são encontradas na distribuição Beta, no entanto a única forma que é considerada nesse estudo é a forma assimétrica à esquerda da média. Esta forma é como um espelhamento da distribuição Exponencial. Os menores valores tem pouca ocorrência, e a medida que os valores vão aumentando, a probabilidade de ocorrência também vai aumentando. Ou seja, os maiores valores possíveis são os mais freqüentemente encontrados.

A distribuição Beta foi utilizada nesse estudo somente na etapa da descoberta da forma da distribuição. Não foi desenvolvido o cálculo dos seus parâmetros.

Segundo GARTHWAITE et al. (2005), as pessoas geralmente ficam mais confortáveis em espressar suas incertezas por meio de termos verbais em vez de termos numéricos. Por conta disso, sempre que possível será dada prioridade aos termos verbais.

Para a descoberta da distribuição de probabilidade foi decidido utilizar uma árvore de decisão (Fig. 3.2) com as características mencionadas de cada distribuição. Cada nó dessa árvore possui uma pergunta relativa a uma particularidade de uma das distribuições citadas. Esta é a fase de extração da informação.

Portanto, foi criado um encadeamento de perguntas sobre as particularidades de cada distribuição, e de acordo com as respostas do especialista a respeito do comportamento da atividade em análise, a folha da árvore apontará para uma das seis

distribuições de probabilidade abordadas nesse estudo. Esta é a fase de ajuste, onde as informações do especialista representam alguma distribuição de probabilidade.

As saídas dos nós são do tipo booleano e fechadas, onde duas opções de respostas são possíveis e são conhecidas, limitando a resposta a um “Sim” ou um “Não”.

O processo interativo entre as perguntas do módulo de elicitação, que estão contidas na árvore de decisão, e o especialista, permite um flexível fluxo do caminho a ser percorrido até a definição de uma distribuição. O especialista pode recomeçar e voltar à pergunta inicial ou voltar à pergunta anterior a qualquer momento do processo de elicitação.

Antes de se chegar a uma conclusão sobre a forma da distribuição de probabilidade, uma pergunta com caráter confirmatório é feita ao especialista, reforçando a pergunta anterior que levou a aquela conclusão. Estas perguntas são as: E1, E2, E3, E4, E5, E6 e E7 contidas na árvore de decisão. Caso ele responda positivamente à pergunta, a distribuição correspondente a aquela folha é selecionada.

Se essa pergunta alertar para um comportamento que não condiz com que o especialista pensa a respeito da atividade, então este tem a opção de recomeçar o processo ou voltar para alguma das perguntas anteriores onde a resposta que ele forneceu tenha sido equivocada.

Após a conclusão obtida através das características da árvore de decisão, outras características da variável são analisadas. Desta vez as características estão relacionadas a atividades que freqüentemente apresentam um determinado comportamento, obtendo a mesma distribuição de freqüência. Lembrando que esta associação da variável analisada com estas características, caso exista, deve ser apontada na etapa de configuração do processo de elicitação.

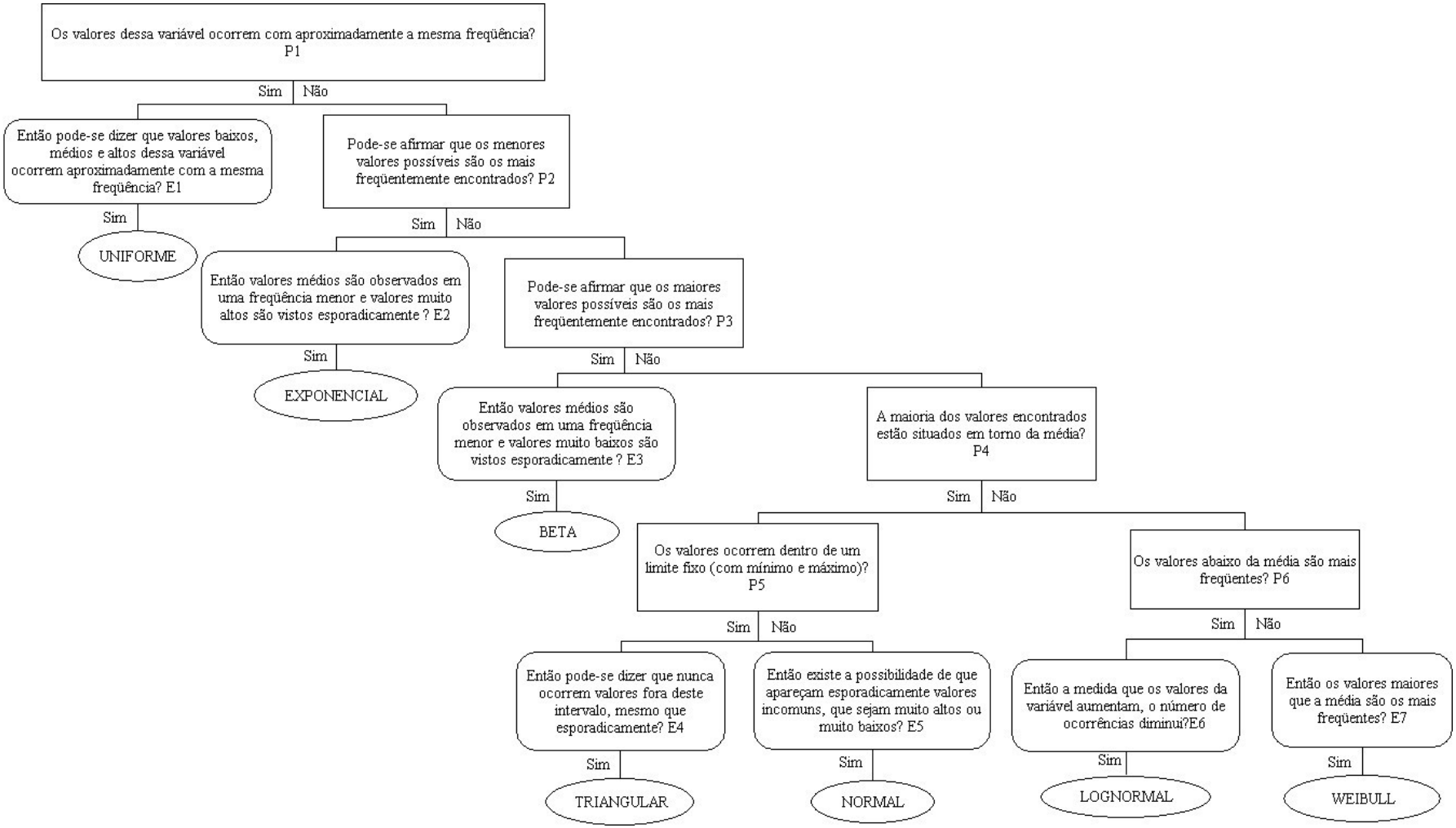

Figura 3.2 – Árvore de decisão do processo de elicitação.

A seguir são apresentadas as características de variáveis encontradas na literatura, seguindo da distribuição de probabilidade que modela seu comportamento:

- Medidas: Medidas de um modo geral como: peso, altura, temperatura e comprimento são observados freqüentemente seguindo uma distribuição Normal.
- Notas e escores: Estas também apresentam comumente um comportamento de uma distribuição Normal.
- Tempo entre dois eventos: Esta característica é bastante comum possuindo uma distribuição Exponencial.
- Distância entre dois pontos: Assim como a anterior, esta característica também é usualmente identificada como seguindo uma distribuição Exponencial.
- Tempo de vida útil (com desgaste): Muitos artigos atestam o fato de que a melhor distribuição para representar este evento é a distribuição Weibull.
- Propriedade de materiais: Vários autores sugerem que a distribuição Lognormal é bastante indicada para este tipo de variável.

Estas são algumas das características existentes e que são adotadas neste trabalho. Pode-se notar que as distribuições Uniforme e Triangular não estão presentes nesta etapa. Isto decorre do fato destas distribuições serem freqüentemente usadas, por serem mais simples.

O objetivo de apresentar estas características citadas é contribuir para que o especialista não indique estas duas distribuições em todos os casos. Para isto são oferecidas algumas informações adicionais sobre as outras distribuições. Porém, o fato de uma variável ter uma das características citadas, não é prova suficiente que a variável tenha esta distribuição.

Por este motivo, caso o especialista indique uma característica que possua distribuição de probabilidade diferente da distribuição elicitada, apenas uma observação a respeito do comportamento da variável é informada ao especialista, cabendo sempre a ele a decisão final de qual distribuição de probabilidade é a mais indicada.

Para auxiliar o especialista nesta escolha, um gráfico em forma de histograma é apresentado contendo a forma da distribuição elicitada. Caso o especialista julgue que aquele gráfico não condiz com o comportamento da atividade que ele está pensando, ele pode voltar atrás e responder novamente as perguntas e seguir assim até encontrar um

gráfico que apresente características semelhantes às da atividade. Podendo também recomeçar todo o processo de elicitação.

Caso a distribuição encontrada pela elicitação seja diferente da distribuição apontada pela característica que ele escolheu no início do processo, dois gráficos, um para cada distribuição, estarão visíveis ao especialista. Facilitando, deste modo, a comparação entre os dois modelos.

O último passo é a fase da adequação: se o especialista não confirmar que algum dos dois gráficos realmente é o modelo mais próximo da realidade, ele pode reiniciar o processo de elicitação, retornando a pergunta inicial.

As outras opções são concordar com o gráfico obtido pela elicitação ou escolher o modelo apresentado como padrão para aquele tipo de variável.

O próximo estágio é o segundo ciclo desta grande etapa, contendo a argüição de valores para a descoberta dos parâmetros desta distribuição escolhida.

3.2 – Procedimento para a descoberta dos parâmetros

Os parâmetros de algumas distribuições são pouco intuitivos ou muito complexos para a percepção natural das pessoas menos familiarizadas com os conceitos estatísticos.

Por isso, a estratégia utilizada é perguntar sobre valores que sejam mais acessíveis a estas pessoas (fase de extração) e a partir desses valores é realizado o calculo dos parâmetros das distribuições (fase de ajuste).

A pesquisa de GARTHWAITE et al. (2005) encontrou alguns trabalhos que concluíram alguns aspectos importantes na difícil tarefa de extrair das pessoas informações subjetivas.

Primeiramente, foi analisada a dificuldade que as pessoas possuem em estimar medidas centrais, como por exemplo, média e mediana. Na pesquisa apontada por este estudo, nas ocasiões em que a distribuição da amostra era simétrica, os entrevistados mostraram uma boa precisão na estimativa da média, mediana e moda.

No entanto, quando a distribuição apresentou uma forte assimetria, para direita ou esquerda, a estimativa dos entrevistados a respeito da mediana e moda permaneceram com uma boa precisão, porém a média sofreu um enviesamento em direção à mediana.

Em outra ocasião neste mesmo estudo foi indicado que estimar valores situados nas caldas de uma distribuição é uma tarefa muito complicada. Isso se dá pelo fato de ser necessária a consideração de valores que são muito incomuns e geralmente as pessoas entrevistadas não têm facilidade em prontamente identificar tais valores.

Tomando como base as indicações anteriores, é mostrado a seguir a descrição das soluções encontradas para o cálculo dos parâmetros para cada distribuição.

3.2.1 – Distribuição Uniforme

A distribuição Uniforme não apresenta nenhum cálculo para a descoberta dos seus parâmetros. É requisitado ao especialista, os mesmos valores que compõem seus parâmetros, que são os valores: *mínimo* e *máximo*.

3.2.2 – Distribuição Triangular

Para a descoberta dos parâmetros da distribuição Triangular, dois valores serão, assim como na distribuição Uniforme, perguntados diretamente ao especialista. Estes parâmetros são: *mínimo* e *máximo*.

Apenas um parâmetro será realmente calculado para esta distribuição, a *moda*. Neste estudo, apenas a distribuição Triangular simétrica foi considerada. Portanto, o cálculo da *moda* se dá por:

$$Moda = \frac{mínimo + máximo}{2} \qquad 3.1$$

Este parâmetro poderia também ser perguntado ao especialista, porém foi convencionado neste estudo que somente duas perguntas poderiam ser feitas ao especialista. Desta forma, como o valor da *moda* pode ser facilmente obtido, de posse dos outros dois parâmetros, por meio da equação 3.1, este parâmetro é calculado e não perguntado ao especialista.

3.2.3 – Distribuição Normal

Os parâmetros da distribuição Normal são μ e σ. O valor μ é um termo usado frequentemente e entendido por todos, porém o σ não é um valor de fácil identificação para as pessoas que não estão acostumadas com os termos estatísticos. De acordo com GARTHWAITE et al. (2005) as pessoas parecem ser deficientes na interpretação do conceito de variância (ou desvio-padrão) e em atribuir-lhe um valor. Devido a isto, valores mais familiares a estas pessoas foram analisados para poder, a partir deles, se determinar os valores mais condizentes com os parâmetros.

Devido ao caráter assintótico da distribuição, valores máximos e mínimos não puderam ser considerados.

Uma alternativa para não ser necessário perguntar diretamente estes parâmetros é solicitar ao especialista que forneça a faixa de valores que compreende 50% da totalidade dos valores em torno de μ.

O especialista deve fornecer dois valores "a" e "b". Sendo que "a" é o limite inferior que irá conter 25% dos valores menores que a média e "b" é o limite superior que irá conter 25% dos valores maiores que μ.

Localizando estes valores acima na tabela da distribuição Normal Padrão, a área que corresponde a 25% do limite inferior ou superior está distante $0{,}67\sigma$ de μ.

Portanto, o valor "a" informado pelo especialista corresponde a (μ - $0{,}67\sigma$) e o valor "b" refere-se a (μ + $0{,}67\sigma$), como pode ser visto na figura 3.3.

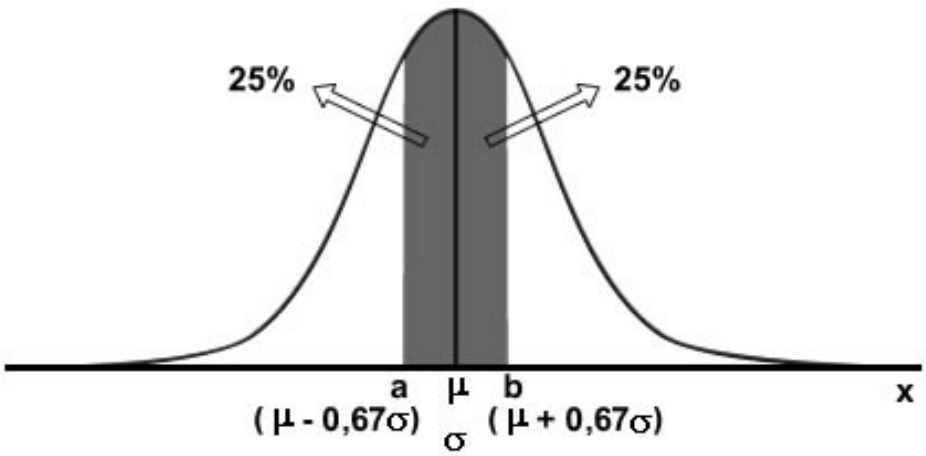

Figura 3.3 – Distribuição Normal: Faixa com 50% dos valores em torno da média

Aplicando estes valores na expressão 2.2, o valor de *X* será o valor "b" indicado pelo especialista. E o valor Z será a equivalência de "b" na tabela da distribuição Normal Padrão, que é o valor 0,67.

Desse modo substituindo Z por 0,67 na expressão 2.2:

$$0{,}67 = \frac{b - \mu}{\sigma} \qquad 3.2$$

Tendo como base a característica de simetria da distribuição, a μ , que é o segundo parâmetro que se deseja encontrar, pode ser calculada da seguinte forma:

$$\mu = \frac{b - a}{2} \qquad 3.3$$

Com o valor da μ adquirido, o cálculo do σ ficará da seguinte maneira:

$$\sigma = \frac{b - \mu}{0{,}67} \qquad 3.4$$

Portanto, a partir de uma pergunta realizada ao especialista a respeito da faixa de valores que se encontra 50% dos valores em torno de μ , pôde-se chegar as expressões 3.3 e 3.4, as quais fornecem os parâmetros reais da distribuição, que são: μ e σ .

3.2.4 – Distribuição Lognormal

Os parâmetros da distribuição Lognormal são: *média* $E(X)$ e *variância* $V(X)$. Estas notações para média e variância serão adotadas para evitar confusões entre a *média* e *variância* da distribuição Normal e a *média* e *variância* da distribuição Lognormal.

Para o cálculo dos parâmetros da distribuição Lognormal é necessário que se tenha os valores dos parâmetros μ e σ de $\ln(X)$, como pode ser observado no quadro 2.4.

É utilizado inicialmente o relacionamento que existe entre a distribuição Normal e a distribuição Lognormal na descoberta dos parâmetros μ e σ, para posterior descoberta dos parâmetros reais (E(X) e V(X)) da distribuição Lognormal.

Com o intuito de descobrir os parâmetros citados, este estudo faz uma associação entre o menor valor da distribuição Normal com o menor valor da distribuição Lognormal. Como apresentado na seção 2.4 desta pesquisa, que tratava da distribuição Normal, o intervalo correspondente a probabilidade de 99,7% de concentração dos valores é compreendido entre $\mu - 3\sigma$ e $\mu + 3\sigma$. Portanto, este limite inferior ($\mu - 3\sigma$) é adotado como uma medida aproximada do menor valor possível.

Para transportar este menor valor da distribuição Normal para a distribuição Lognormal é necessário elevar o número de *Euller* (*e*) a potência do menor valor da distribuição Normal:

$$e^{(\mu - 3\sigma)} \qquad 3.5$$

Outra associação é feita entre o valor referente a *moda* da distribuição Normal com o valor da *moda* da distribuição Lognormal. O valor para a medida da *moda* encontrada no quadro 2.4 é:

$$e^{(\mu - \sigma^2)} \qquad 3.6$$

Estas associações realizadas anteriormente podem ser visualizadas na figura 3.4.

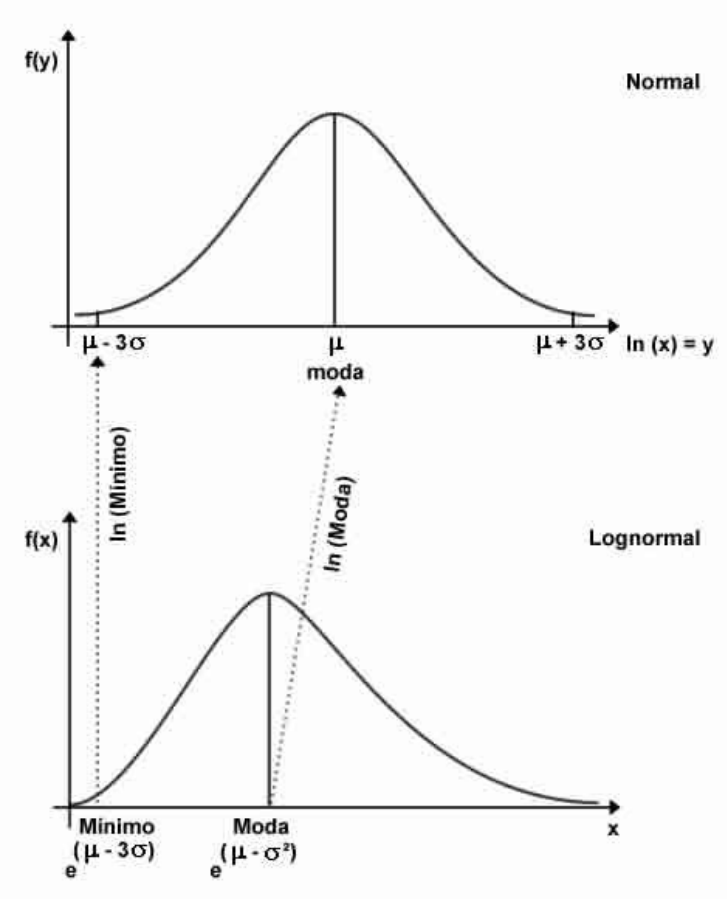

Figura 3.4 – Mapeamento de valores da distribuição Lognormal para a Normal.

Portanto ao especialista (especialista) é perguntado o menor valor e a *moda* da distribuição Lognormal. Então:

$$\text{Mínimo} = e^{\mu - 3\sigma} \qquad 3.7$$

Como se deseja saber primeiramente os valores de μ e σ, esta igualdade é transportada para a distribuição Normal. Para isto é necessário aplicar o logaritmo neperiano em ambos os valores.

$$\ln(\text{Mínimo}) = \mu - 3\sigma \qquad 3.8$$

O mesmo raciocínio aplicado anteriormente é utilizado para o valor da *moda*. Desta forma:

$$\ln(\text{Moda}) = \mu - \sigma^2 \qquad 3.9$$

Isolando a μ da expressão 3.9:

$$\mu = \ln(\text{Moda}) + \sigma^2 \qquad 3.10$$

Substituindo o valor de μ da expressão 3.8 pelo valor da expressão 3.10:

$$\ln(\text{Moda}) + \sigma^2 = \ln(\text{Mínimo}) + 3\sigma \qquad 3.11$$

$$\sigma^2 - 3\sigma + \ln(\text{Moda}) - \ln(\text{Mínimo}) = 0 \qquad 3.12$$

$$\sigma = \frac{3 \pm \sqrt{9 - 4(\ln Moda - \ln Mínimo)}}{2} \qquad 3.13$$

O único resultado válido é obtido utilizando o sinal negativo na expressão 3.13.

Após ter sido calculado o σ, pode-se então descobrir o valor da μ por meio da expressão 3.10.

De posse dos parâmetros da distribuição Normal, basta aplicá-los na fórmula da média e variância da distribuição Lognormal apresentada no quadro 2.4.

3.2.5 – Distribuição Exponencial

A distribuição Exponencial, como apresentado na seção 2.6, apresenta λ como parâmetro. É perguntado ao especialista o valor referente à *mediana* para se encontrar λ, utilizando para isto a expressão 3.14, que é a relação explicitada no quadro 2.5 entre este parâmetro e a *mediana*.

$$Mediana = \frac{\ln 2}{\lambda} \qquad 3.14$$

Adicionalmente a *mediana*, outro valor é requisitado ao especialista. Este valor é referente ao valor *mínimo*. Como exposto anteriormente, nenhum outro valor é necessário para calcular λ além da *mediana*. Entretanto, para a FGVA da distribuição Exponencial há a necessidade de saber o valor *mínimo*, caso contrário, valores próximos de zero são sempre adotados como referencia para o menor valor gerado por simulação.

Utilizando um valor mínimo condizente com o comportamento da atividade em análise, a geração dos valores, por simulação, representará satisfatoriamente a atividade real.

3.2.6 – Distribuição Weibull

A solução proposta para a descoberta dos parâmetros α e β da distribuição Weibull é dividida em duas etapas: a primeira deriva da fda e a segunda tem sua origem no valor da *moda*. Nesta solução, as informações solicitadas ao especialista referem-se aos valores *moda* e *máximo*.

A fda de x se refere à probabilidade de um valor ser igual ou abaixo de x. Para este valor de x é utilizado o valor *máximo* que foi informado pelo especialista. Desta forma, a fda F(x) de uma variável contínua *X* fornece, para qualquer valor de x, a probabilidade $P(X \leq x)$ (DEVORE, 2006). Como x é o valor *máximo*, a probabilidade $F(Max) = P(X \leq Max) = 1$.

Portanto, a primeira etapa desta solução origina desta igualdade. A seguir é descrita passo a passo a dedução matemática desta solução.

Etapa 1 – Solução a partir da fda
F(Max) = 1

$$1 - e^{-\left(\frac{Max}{\beta}\right)^{\alpha}} = 1 \qquad 3.15$$

$$e^{-\left(\frac{Max}{\beta}\right)^{\alpha}} = 0 \qquad 3.16$$

Como não existe ln de zero, então será utilizado um valor próximo de zero.

$$e^{-\left(\frac{Max}{\beta}\right)^{\alpha}} = 0{,}0001 \qquad 3.17$$

$$\ln\left(e^{-\left(\frac{Max}{\beta}\right)^{\alpha}} \right) = \ln 0{,}0001 \tag{3.18}$$

$$-\left(\frac{Max}{\beta}\right)^{\alpha} . \ln e = -9{,}21034 \tag{3.19}$$

Como ln e = 1, então:

$$\left(\frac{Max}{\beta}\right)^{\alpha} = 9{,}21034 \tag{3.20}$$

$$\ln\left(\frac{Max}{\beta}\right)^{\alpha} = \ln 9{,}21034 \tag{3.21}$$

$$\alpha\left(\ln Max - \ln\beta\right) = \ln 9{,}21034 \tag{3.22}$$

$$\ln Max - \ln\beta = \frac{2{,}2203268}{\alpha} \tag{3.23}$$

$$ln\,\beta = \left(\ln Max - \frac{2{,}2203268}{\alpha}\right) \tag{3.24}$$

Etapa 2 – Solução a partir da moda

$$Moda = \beta\left(\frac{\alpha - 1}{\alpha}\right)^{\frac{1}{\alpha}} \tag{3.25}$$

$$ln\ Moda = ln\,\beta + ln\left(\frac{\alpha - 1}{\alpha}\right)^{\frac{1}{\alpha}} \tag{3.26}$$

$$ln\ Moda = ln\,\beta + \frac{1}{\alpha}\left[\ln(\alpha - 1) - \ln\alpha\right] \tag{3.27}$$

$$ln\,\beta \;=\; ln\,Moda - \frac{1}{\alpha}\left[\ln(\alpha - 1) - \ln\alpha\right] \tag{3.28}$$

Igualando a expressão 3.24 com expressão 3.28:

$$\left(\ln Max - \frac{2{,}2203268}{\alpha}\right) \;=\; ln\,Moda - \frac{1}{\alpha}\left[\ln(\alpha - 1) - \ln\alpha\right] \tag{3.29}$$

$$ln\,Max - ln\,Moda = 2{,}2203268 \cdot \frac{1}{\alpha} - \frac{1}{\alpha}\left[\ln(\alpha - 1) - \ln\alpha\right] \tag{3.30}$$

$$ln\,Max - ln\,Moda = \frac{1}{\alpha}\left\{2{,}2203268 - \left[\ln(\alpha - 1) - \ln\alpha\right]\right\} \tag{3.31}$$

$$(ln\,Max - ln\,Moda.\alpha \;=\; 2{,}2203268 - \left[\ln(\alpha - 1) - \ln\alpha\right] \tag{3.32}$$

$$(ln\,Max - ln\,Moda)\,\alpha \;=\; 2{,}2203268 - \ln\left[\frac{(\alpha - 1)}{\alpha}\right] \tag{3.33}$$

$$\ln\left[\frac{(\alpha - 1)}{\alpha}\right] \;=\; 2{,}2203268 - (ln\,Max - ln\,Moda).\alpha \tag{3.34}$$

$$e^{\left\{2{,}2203268 - \left[(\ln Max - \ln Moda)\alpha\right]\right\}} \;=\; \frac{\alpha - 1}{\alpha} \tag{3.35}$$

$$\frac{e^{2{,}2203268}}{e^{(\ln Max - \ln Moda)\alpha}} = \frac{\alpha - 1}{\alpha} \tag{3.36}$$

$$e^{2{,}2203268} = e^{(\ln Max - \ln Moda).\alpha} \cdot \left(\frac{\alpha - 1}{\alpha}\right) \tag{3.37}$$

$$e^{(\ln Max - \ln Moda).\alpha} \cdot \left(\frac{\alpha - 1}{\alpha}\right) = 9{,}21034 \qquad 3.38$$

Com a expressão 3.38 se determina o parâmetro de escala α , atribuindo vários valores para α até que se encontre um valor que satisfaça esta igualdade. Como α tem que ser maior que 4 para a distribuição Weibull ter a forma assimétrica à esquerda, o primeiro valor atribuído a ele é exatamente 4.

Utilizando a expressão 3.20 para calcular β :

$$\left(\frac{Max}{\beta}\right)^{\alpha} = 9{,}21034 \qquad 3.39$$

$$\frac{Max^{\alpha}}{\beta^{\alpha}} = 9{,}21034 \qquad 3.40$$

$$\beta^{\alpha} = \frac{Max^{\alpha}}{9{,}21034} \qquad 3.41$$

$$\alpha . \ln \beta = \ln\left(\frac{Max^{\alpha}}{9{,}21034}\right) \qquad 3.42$$

$$\ln \beta = \frac{1}{\alpha} . \ln\left(\frac{Max^{\alpha}}{9{,}21034}\right) \qquad 3.43$$

$$\beta = e^{\left[\frac{1}{\alpha} . \ln\left(\frac{Max^{\alpha}}{9{,}21034}\right)\right]} \qquad 3.44$$

Depois de α calculado, utiliza-se a expressão 3.44 para a obtenção de β . Com isto, a partir dos valores da moda e do máximo, informados pelo especialista, pode-se calcular os parâmetros α e β .

Como dito anteriormente, há a possibilidade da existência de um terceiro parâmetro, o δ . Porém, com a solução proposta, no momento da simulação de vetores

de dados baseados na distribuição Weibull com os valores de α e β calculados, este terceiro parâmetro é calculado automaticamente.

O quadro 3.1 apresenta um resumo dos valores de entrada, que representam os dados informados pelo especialista e os parâmetros de cada distribuição apresentada nesta dissertação.

Quadro 3.1 – Valores de entrada x Parâmetros

Distribuição	Valor de entrada	Parâmetros
Uniforme	mínimo, máximo	Mínimo, máximo
Triangular	mínimo, máximo	Mínimo, moda e máximo
Normal	faixa de 50% em torno da média	μ e σ
Lognormal	mínimo, moda	E(X) e V(X)
Exponencial	mínimo, mediana	λ
Weibull	moda, máximo	α e β

Utilizando novamente um artifício visual para uma avaliação da distribuição indicada, um novo histograma é apresentado. Mas dessa vez ele é gerado por um programa gerador de números aleatórios baseado na distribuição de probabilidade encontrada juntamente com os parâmetros calculados.

Além de apresentar a forma da distribuição, o histograma apresenta também a escala de valores. O especialista pode observar no histograma, em qual valor está o maior número de observações, quais são os valores extremos, os valores mínimos e máximos, dentre outras informações.

Novamente é dada ao especialista a possibilidade de recomeçar o processo de elicitação, caso a escolha da distribuição não esteja adequada. Desta vez, o recomeço é desde a etapa de configuração.

Se o especialista confirmar que aquele gráfico apresenta características semelhantes às da variável analisada, o último procedimento que permite uma observação geral de todo o processo de elicitação é acionado.

Neste procedimento, outra geração de números aleatórios baseada na distribuição de probabilidade escolhida é efetuada. Mas desta vez, ocorrem cinco simulações com os mesmos parâmetros, gerando cinco vetores de dados. Este processo é realizado para dar possibilidade da análise de múltiplas visões. O especialista deve escolher qual delas é a mais adequada para a variável. Dependendo da visão escolhida, os parâmetros da distribuição terão valores diferentes.

Estas visões se baseiam no conceito da visão otimista, pessimista e realista. Na visão otimista são considerados os menores valores de cada posição dos cinco vetores gerados. A média dos valores é caracterizada na visão realista. E a visão pessimista é composta pelos valores máximos de cada posição do vetor.

Porém, como estes conceitos podem ter interpretações diferentes, já que variáveis de diversas características podem ser utilizadas neste processo, novas nomenclaturas foram utilizadas.

Para exemplificar as mudanças de interpretações utilizando os nomes: otimista, pessimista e realista, basta pensar em duas atividades, uma relacionada com tempo de execução de um determinado serviço e uma outra com nota.

Uma visão otimista está relacionada com a conclusão deste serviço no menor tempo possível. No entanto, uma visão otimista, no caso de notas atribuídas a um serviço, seria o maior valor possível. Pois quanto maior a nota, melhor é o serviço.

Então para uma atividade, o menor valor é a visão otimista, enquanto que para o outro é justamente o oposto, seria uma visão pessimista.

Por causa desde tipo de confusão que pode ocorrer com esta nomenclatura, as visões passaram a ser denominadas de valores: mínimos, médios e máximos.

Cada visão é representada por um gráfico separadamente e por um único gráfico contendo todas as três juntas. Essa unificação permite uma comparação melhor entre elas, ajudando na escolha.

Como o processo de elicitação é algo inerentemente iterativo e interativo, novamente o especialista pode voltar na etapa anterior ou até mesmo reiniciar o processo de elicitação, fazendo a adequação. Ou pode finalizar o processo, salvando este cenário elicitado que contém a distribuição de probabilidade juntamente com seus parâmetros.

Capítulo 4

Implementação do Módulo de Elicitação

Este capítulo especifica alguns detalhes do software desenvolvido, chamado de Módulo de Elicitação. Para isto foi utilizado o diagrama de seqüência da UML (Unified Modeling Language). Este diagrama é voltado a descrever objetos interagindo, onde o tempo decorre de cima para baixo (SILVA, 2007). Após esta documentação, são apresentadas as interfaces do Módulo, bem como a explanação de cada item nelas contido.

4.1 – Especificação Formal do Software

Esta seção descreve formalmente, por meio do diagrama de seqüência, os fluxos que podem ser percorridos dentro do Módulo de Elicitação. Estes fluxos são formados pela seqüência cronológica de passos, desde o início do Módulo até um ponto final. Este ponto final não necessariamente atinge o final do Módulo com a devida resposta. Outras situações são previstas devido a interação que existe entre o Módulo e o especialista que o está usando.

Inicialmente, na seção 4.1.1., é ilustrado o fluxo denominado de fluxo padrão, onde o especialista, também conhecido apenas como usuário do Módulo, percorre um trajeto linear no Módulo, desde o início até o final, obtendo o resultado desejado.

Posteriormente, nas demais seções da Especificação Formal do Software, são apresentadas o restante dos fluxos alternativos do sistema.

Existe, nos diagramas, uma ação que é executada ou pelo usuário ou pelo Módulo de Elicitação. Esta ação leva um nome no diagrama e vem acompanhada de uma seta, que indica quem está participando da ação.

Para facilitar o entendimento de como se processa a ação, o Quadro 4.1 possui o nome de todas as ações que aparecem nos diagramas, bem como uma descrição desta ação.

Quadro 4.1 – Descrição dos eventos do diagrama de seqüência.

Evento	Descrição
MostraTelaApresentacao	Apresenta a tela Apresentacao.
ConfirmElicitar	O usuário confirma que deseja elicitar a distribuição de probabilidade de uma variável, clicando no botão “Elicitar”.
MostraTelaCadastro	Apresenta a tela Cadastro para o usuário.
InformaDadosCadastrais	O usuário informa os dados dos seguintes campos: Usuário, Domínio, Descrição, Nome da atividade, Rótulo e Característica da variável.
ArmazenaDadosCadastrais	Armazena os dados informados pelo usuário em variáveis do sistema.
MostraTelaPreProcesso1	Apresenta a tela PreProcesso 1.
InformaVarQT	O usuário informa que a variável é quantitativa.
InformaVarQL	O usuário informa que a variável é qualitativa
MostraTelaPreProcesso2	Apresenta a tela PreProcesso 2.
InformaVarCont	O usuário informa que a variável é contínua.
InformaVarDisc	O usuário informa que a variável é discreta.
MostraTelaPerguntas	Apresenta a tela Perguntas.
EncontraPerguntaIni	Busca a primeira pergunta da árvore de decisão.
RetornaPerguntaIni	Apresenta a primeira pergunta na tela Perguntas.
RespSimPergIni	O usuário responde “sim” para a pergunta inicial.
EncontraPerguntaP1	Busca a pergunta P1 da árvore de decisão.
RetornaPerguntaP1	Apresenta a pergunta P1 na tela Perguntas.
AtingirLimite	Este evento verifica se ainda existem perguntas da árvore de decisão para serem apresentadas ao usuário.
SimAtingiuLimite	Atingiu o limite de perguntas, ou seja, não há mais perguntas a serem apresentadas ao usuário.
NaoAtingiuLimite	Ainda não atingiu o limite de perguntas da árvore de decisão. Portanto ainda há perguntas a serem apresentadas ao usuário.

(continua na próxima página)

(continuação)

VerificaObservacao	Verifica se o usuário informou alguma característica de uma distribuição na tela Cadasto. Caso tenha informado, verifica se a distribuição elicitada é a mesma distribuição que apresenta a característica informada.
MostrarTelaFigDistSemObs	Apresenta a tela FigDist com a figura da distribuição encontrada. Não apresenta observação para a distribuição.
MostrarTelaFigDistComObs	Apresenta a tela FigDist com a figura da distribuição encontrada além de apresentar também uma figura da distribuição que possui a característica informada na tela cadastrar.
ConfirmaDist	O usuário confirma que a figura apresentada representa o comportamento dos dados clicando o botão “sim”.
Recomecar	O usuário volta para a tela Cadastrar e recomeça o processo de elicitação.
EncontraDistSelec	Encontra no objeto Lógica, qual distribuição de probabilidade foi selecionada no programa.
InformaValores	O usuário informa os valores requeridos pela tela Parâmetros 1.
RequerCalcParam	A tela Parâmetros envia os valores fornecidos pelo usuário para o objeto Lógica. Este irá calcular os parâmetros da distribuição encontrada e irá gerar um vetor de dados por meio de simulação.
MostraTelaParametros2	Apresenta a tela Parâmetros 2 com o gráfico da distribuição gerada pelo vetor de dados simulado.
RequerCalcVisoes	A tela Parâmetros 2 chama o objeto Lógica. Este irá simular cinco vetores de mil posições cada e a partir desses vetores ele irá calcular três vetores: 1– com os mínimos; 2 – com a média e 3 – com os máximos.

(continua na próxima página)

(continuação)

RetornaSimulaDist	Retorna o vetor de dados simulado.
MostraTelaParametros 1	Apresenta a tela Parâmetros 1.
ChamaObjetoConclusao	A tela Parâmetros chama o objeto Conclusão. A partir desse momento serão executados os métodos do objeto Conclusão. Porém a tela ainda não será apresentada ao usuário.
RetornaDadosSimul	O objeto Lógica informa os valores dos vetores.
GerarGraficos	O objeto Conclusão gera quatro gráficos utilizando os vetores informados pelo objeto Lógica.
MostraTelaConclusao	Apresenta a tela Conclusão.
SalvaElicitacao	O usuário salva algumas informações do processo de elicitação, são elas: As informações cadastrais, a forma e os parâmetros da distribuição elicitada.
NaoConfirmaDist	O usuário não confirma que a figura apresentada representa o comportamento dos dados clicando o botão “não”.
MostraTelaNaoConfirma	Apresenta a tela NaoConfirma.
Cancelar	O usuário cancela a opção de não confirmar a distribuição e volta para a tela Parâmetros 2
RespVoltar	O usuário volta para a tela anterior.
ConfirmaDistObs	O usuário confirma que a distribuição apresentada pela observação representa o comportamento dos dados.

Os eventos: EncontraPerguntaP1, RetornaPerguntaP1, RespSimPergP1 e RespNaoPergP1 possuem o mesmo funcionamento para as demais perguntas da árvore de decisão: P2, P3, P4, E1, E2, etc.

O evento MostraTelaFigUniforme possui o mesmo funcionamento para os eventos: MostraTelaFigTriangular, MostraTelaFigNormal, MostraTelaFigLognormal, MostraTelaFigExponencial e MostraTelaFigWeibull.

Devido o fato do diagrama de seqüência ser bastante extenso, duas medidas foram tomadas. O diagrama foi dividido em três partes e a opção de sair do Módulo foi ocultada dos diagramas, porém ela é encontrada em todas as telas.

4.1.1 – Fluxo Padrão

A primeira parte do diagrama do fluxo padrão faz referência ao início do Módulo, que é formada por uma tela de apresentação (Apresentação), um cadastro do usuário (Cadastro) e uma tela com perguntas a respeito da natureza da variável sob análise (PreProcesso 1 e 2).

A tela de apresentação é um documento introdutório que contém o objetivo do estudo e estabelece a sua relevância, de acordo com as recomendações de AYYUB (2001).

As demais telas iniciais têm como objetivo sintonizar o usuário com o processo de elicitação, o forçando a restringir seu foco de atenção em uma determinada variável.

A Fig. 4.1 ilustra a primeira parte do diagrama de seqüência do fluxo padrão.

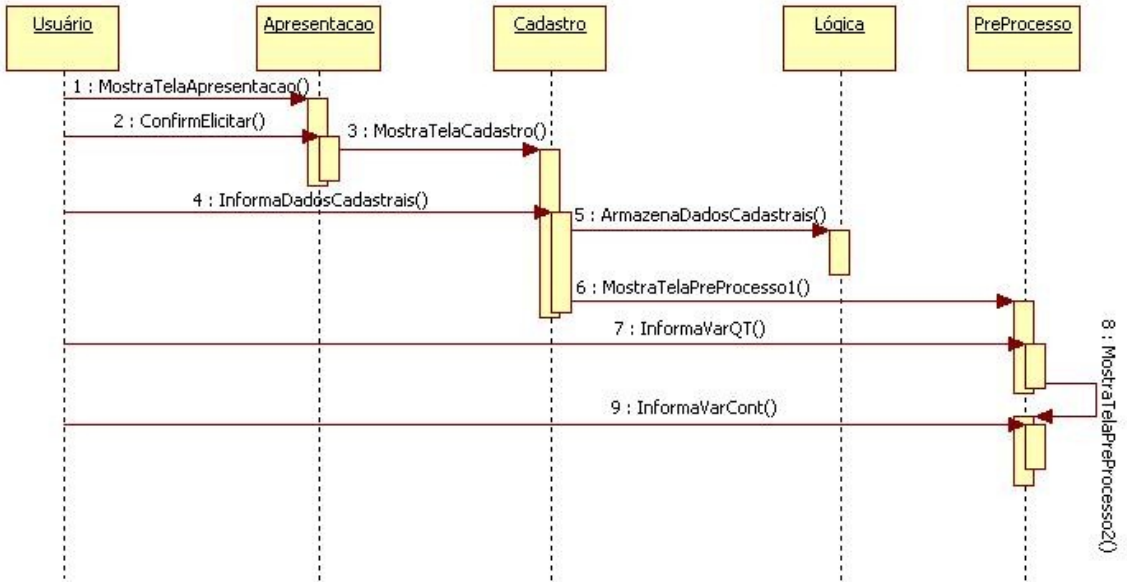

Figura 4.1 – Primeira parte do diagrama de seqüência do fluxo padrão.

A segunda parte do diagrama além de possuir a ligação entre a primeira parte e a segunda, se concentra na elicitação da forma da distribuição, através da tela Perguntas. Esta parte do diagrama mostra como se dá o processo de interação entre o usuário, a tela Perguntas, que apresenta as perguntas da árvore de decisão ao usuário e o objeto Lógica, que é responsável por conter toda a lógica do Módulo.

É este objeto Lógica que, a partir da resposta do usuário a uma dada pergunta, consegue avaliar qual será a próxima pergunta a ser realizada, caso ainda haja perguntas.

Após o término das perguntas, a próxima tela a ser apresentada é a tela FigDist, que mostra graficamente o formato da distribuição de probabilidade em forma de histograma.

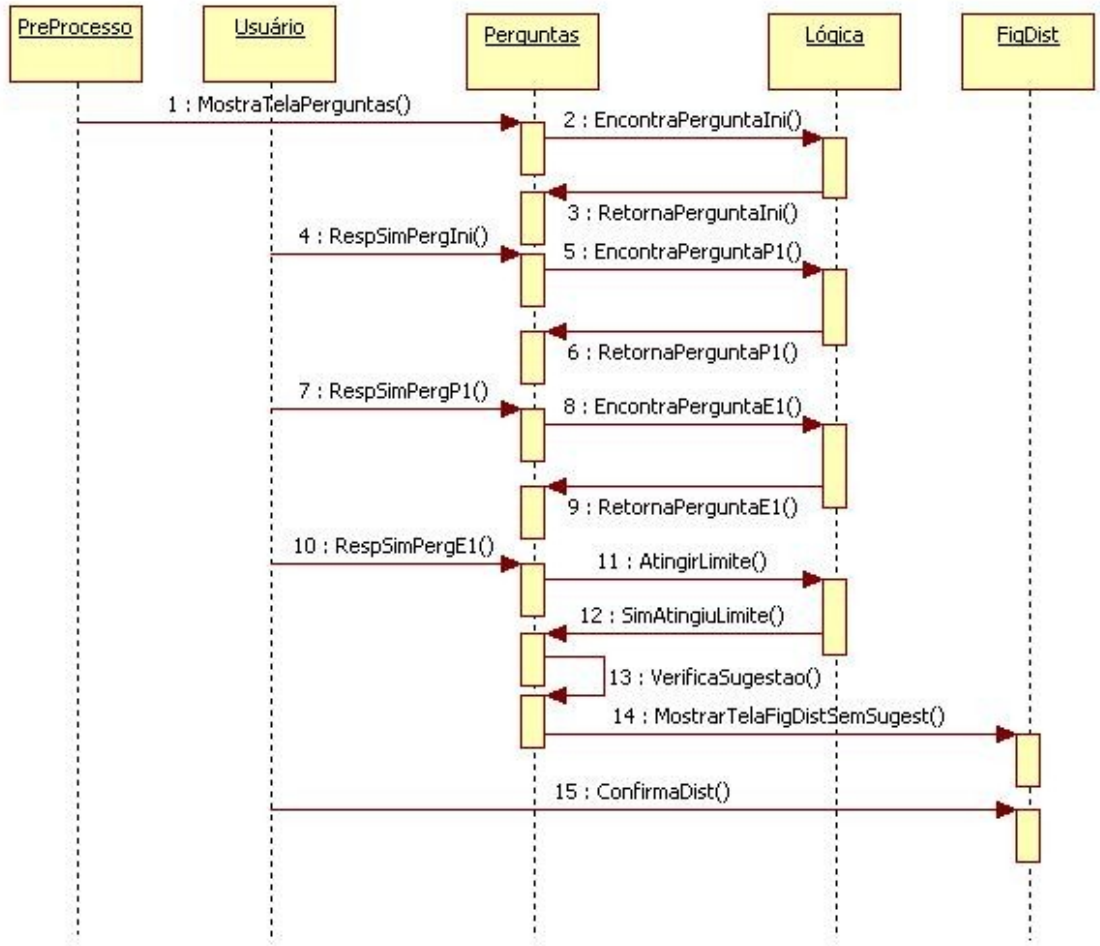

Figura 4.2 – Segunda parte do diagrama de seqüência do fluxo padrão.

Com a confirmação do usuário que aquela forma é a desejada, a terceira parte do diagrama é então requisitada.

Desta forma, esta última parte do diagrama tem como principal foco o cálculo dos parâmetros da distribuição encontrada na etapa anterior. A tela Parâmetros 1 é responsável por obter do usuário os valores necessários para o cálculo dos parâmetros e a tela Parâmetros 2 apresenta um histograma gerado por uma simulação da distribuição encontrada juntamente com os parâmetros calculados.

Assim como na etapa anterior, o objeto Lógica é o responsável pela lógica do Módulo, deixando a parte de relacionamento com o usuário para as outras telas.

Este objeto Lógica que possui as fórmulas e cálculos necessários para a transformação dos valores fornecidos pelo usuário para os valores dos parâmetros das distribuições.

Adicionalmente ao propósito do cálculo dos parâmetros, também consta neste diagrama a fase final do processo de elicitação, que se encontra na tela Conclusão. Esta tela possui as informações finais a respeito da forma e parâmetros da distribuição.

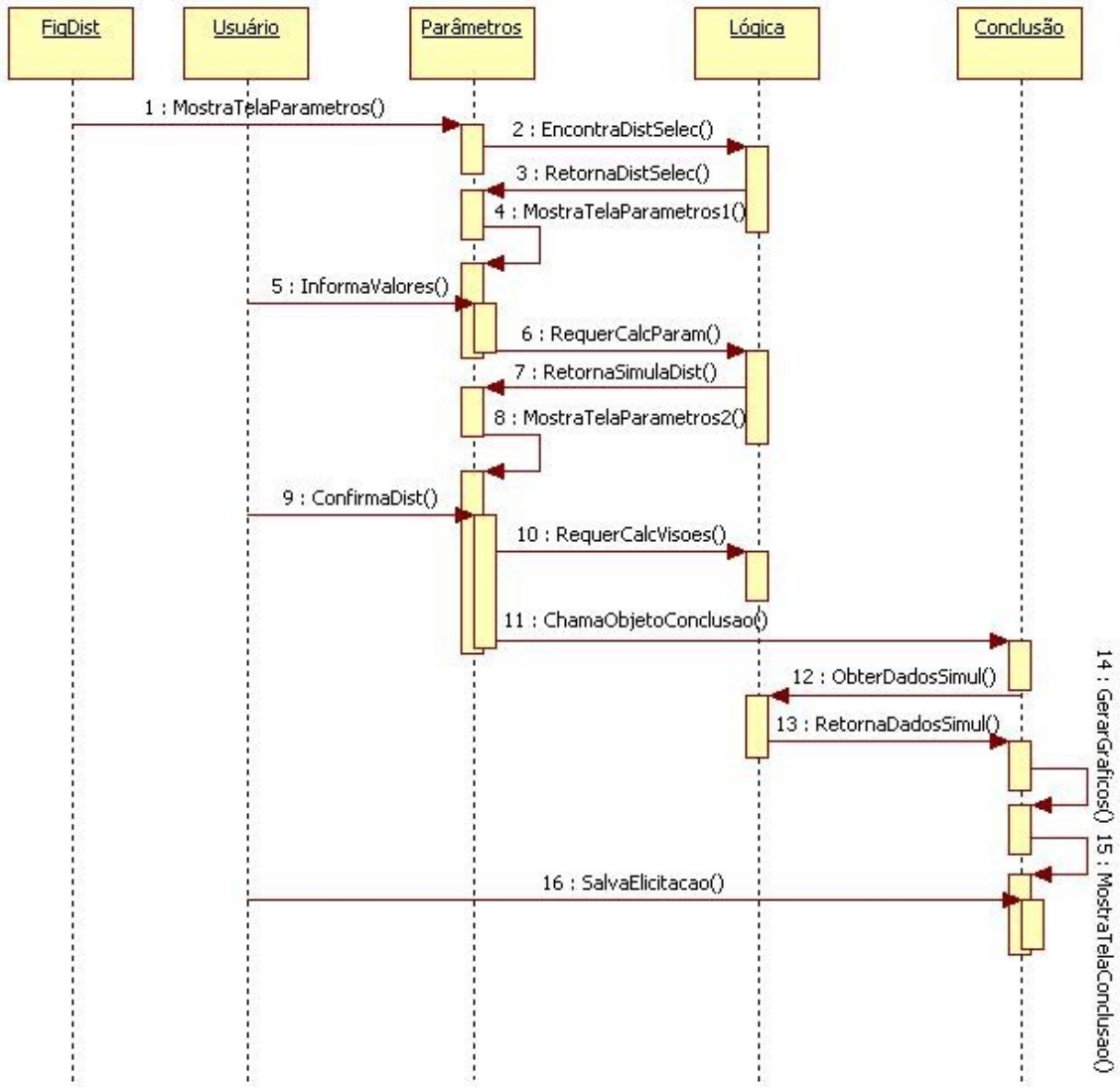

Figura 4.3 – Terceira parte do diagrama de seqüência do fluxo padrão.

A seguir são apresentados os fluxos alternativos ao fluxo padrão, totalizando oito possibilidades de fluxo do Módulo de Elicitação.

4.1.2 – Fluxo Alternativo 1

O fluxo alternativo 1 difere do fluxo padrão na segunda parte do diagrama. A primeira e a terceira parte do diagrama são exatamente iguais ao do fluxo padrão.

No fluxo padrão não era apresentado ao usuário nenhuma observação a respeito da característica da distribuição na tela Cadastro.

Já neste fluxo alternativo, existe uma observação à cerca da distribuição elicitada. De acordo com a literatura atual, variáveis com a característica informada possuem uma distribuição diferente da elicitada.

No entanto se trata apenas de uma observação, cabe ao usuário analisar e decidir se continua com esta distribuição ou não. Neste fluxo, apesar da observação, o usuário decide que a distribuição elicitada, apresentada a ele, é a que melhor representa a variável sob análise. A Fig. 4.4 apresenta a segunda parte do diagrama de seqüência para o fluxo alternativo 1.

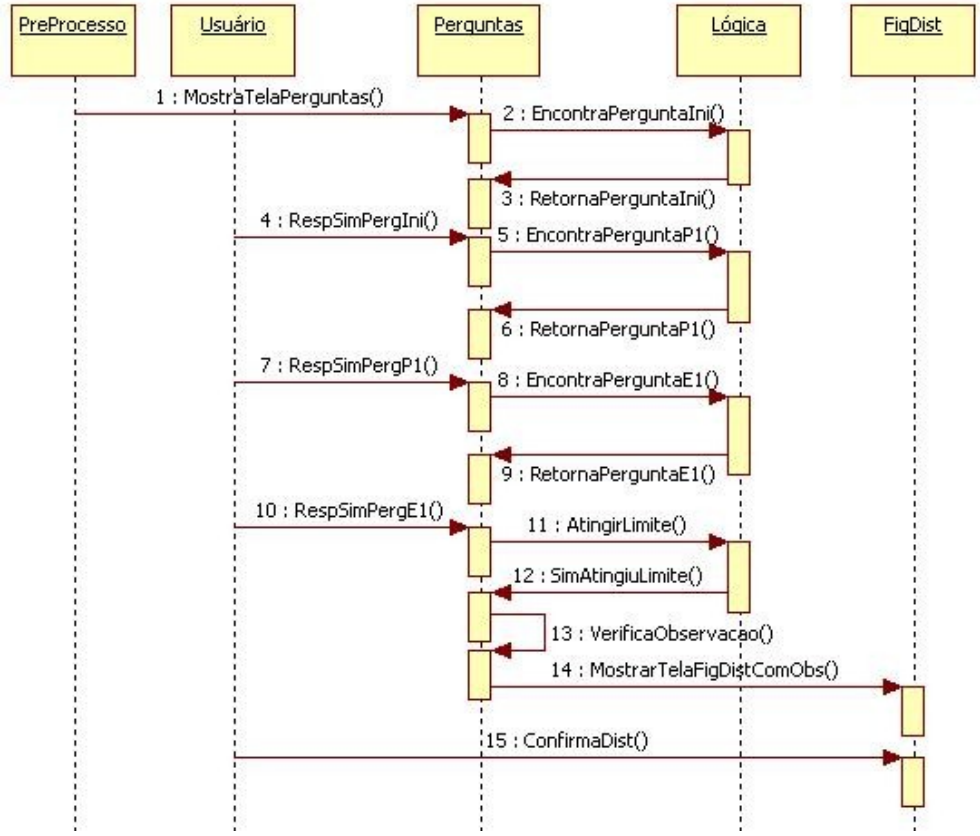

Figura 4.4 – Segunda parte do diagrama de seqüência do fluxo alternativo 1.

4.1.3 – Fluxo Alternativo 2

Similarmente ao fluxo alternativo 1, no fluxo alternativo 2 há uma observação referente a distribuição de probabilidade comumente encontrada para a característica apontada na tela cadastro.

Neste caso, o usuário em uma reflexão maior percebe que a distribuição indicada na observação realmente representa melhor o comportamento da variável do que aquela distribuição que foi encontrada por meio do processo de elicitação.

Desta forma, o usuário não confirma que a distribuição elicitada representa a variável. Neste momento são dadas ao usuário duas alternativas. A primeira alternativa, que é representada pela Fig. 4.5, o usuário decide optar pela distribuição indicada na observação e seguir adiante o processo de elicitação. A segunda alternativa é apresentada na seção 4.1.4.

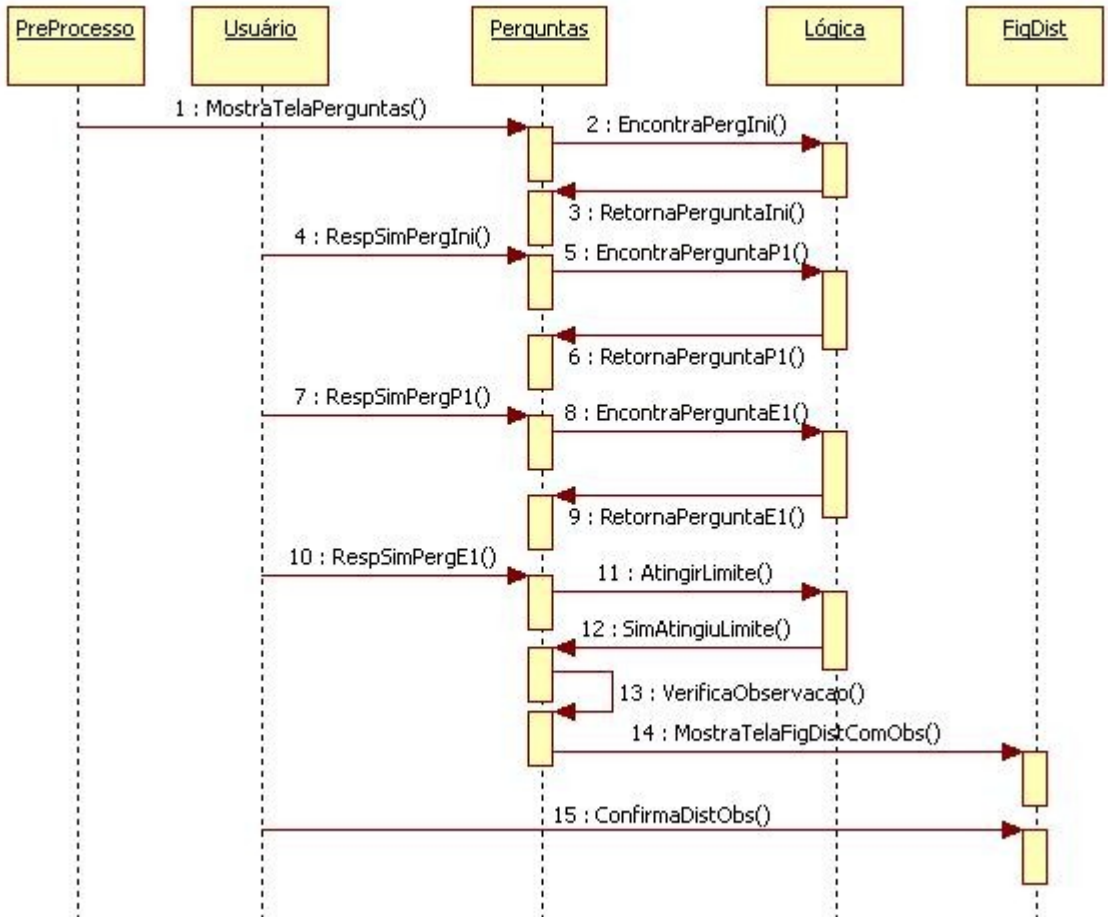

Figura 4.5 – Segunda parte do diagrama de seqüência do fluxo alternativo 2.

4.1.4 – Fluxo Alternativo 3

A segunda alternativa para a situação apresentada na seção 4.1.3 é o usuário recomeçar o processo de elicitação desde a pergunta inicial da árvore de decisão.

Isto pode ser necessário caro o usuário fique em dúvida a respeito da diferença entre o modelo de distribuição encontrado pelo processo de elicitação e o modelo de distribuição apresentado como sendo o mais encontrado na literatura para aquele tipo de variável especifico.

A Fig. 4.6 apresenta o diagrama de seqüência para este caso. Assim como no fluxo de alternativa 2, a primeira e terceira parte do diagrama de seqüência são iguais ao fluxo padrão.

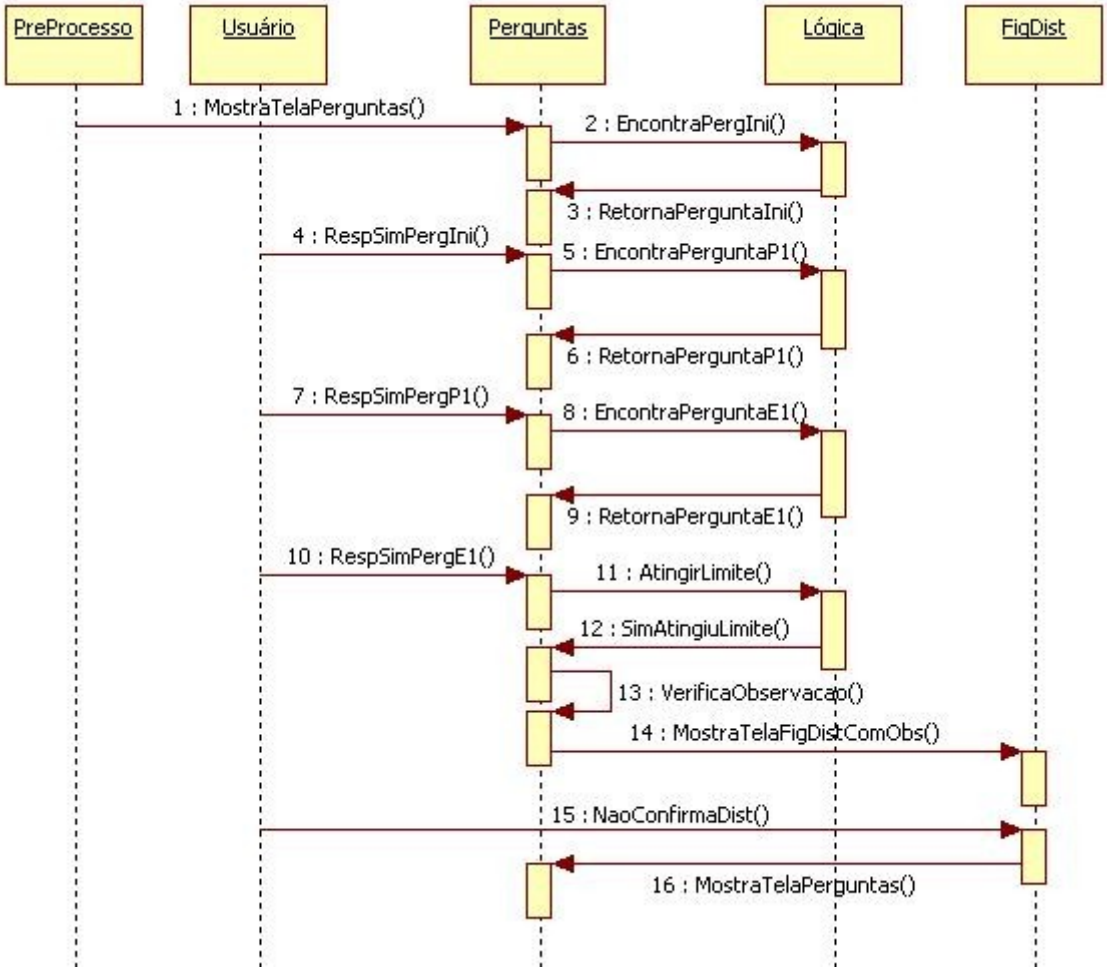

Figura 4.6 – Segunda parte do diagrama de seqüência do fluxo alternativo 3.

4.1.5 – Fluxo Alternativo 4

O fluxo alternativo 4, diferentemente dos fluxos alternativos anteriores, apresenta na terceira parte do diagrama de seqüência sua diferenciação do fluxo padrão.

Na tela Parâmetros 2 há a necessidade de se confirmar se a distribuição encontrada é o melhor modelo para a variável. Entretanto, dessa vez é possível visualizar a distribuição gerada por simulação de acordo com os valores informados pelo usuário.

Logo, esta figura possui uma escala que deve se aproximar do encontrado no ambiente real. Caso esta verificação seja inverídica, algumas possibilidades são oferecidas ao usuário.

Após ser mostrada a tela NaoConfirma para o usuário, este pode recomeçar o processo de elicitação desde o início, voltando para a tela Cadastro. Esta possibilidade é visualizada na Fig. 4.7. Ou pode cancelar a operação e retornar para a tela Parâmetros 2.

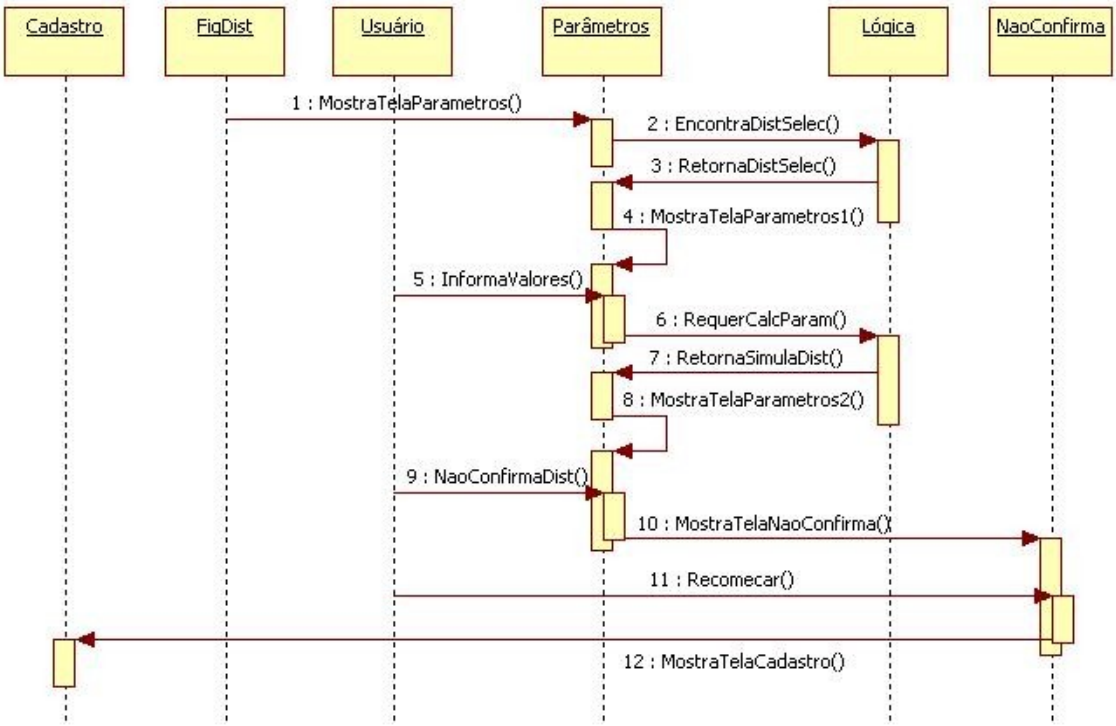

Figura 4.7 – Terceira parte do diagrama de seqüência do fluxo alternativo 4.

4.1.6 – Fluxo Alternativo 5

Outra possibilidade que é dada ao usuário caso este verifique que a escala de valores não condiz com a realidade, apesar da distribuição representar o comportamento da variável, é retornar à tela Parâmetros 1 e redefinir os valores informados anteriormente.

Após um novo refinamento dos valores, tendo como base a simulação anterior, o usuário pode seguir adiante no processo de elicitação ou continuar neste processo de ajuste fino até encontrar a escala correta de valores para sua variável. A Fig. 4.8 apresenta o diagrama de seqüência referente a este processo.

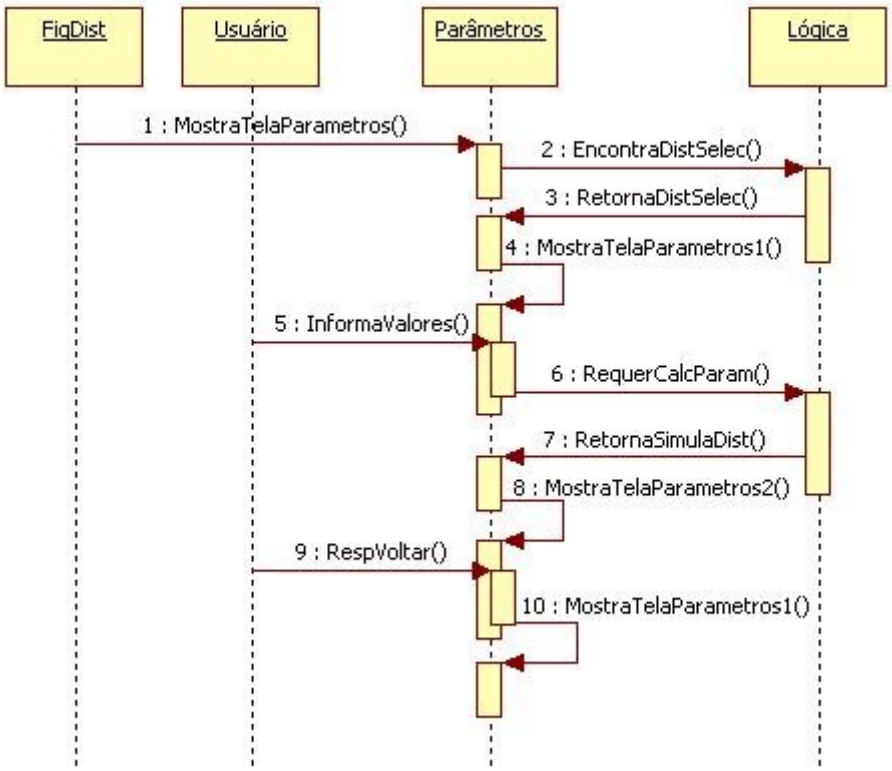

Figura 4.8 – Terceira parte do diagrama de seqüência do fluxo alternativo 5.

4.1.7 – Fluxo Alternativo 6

Os últimos dois fluxos alternativos estão relacionados a primeira parte do diagrama de seqüência. Durante a fase inicial é requisitado ao usuário que responda a duas perguntas referentes a classificação da variável em análise.

A primeira questão é se a variável é qualitativa ou quantitativa. Como o Módulo de Elicitação se restringe apenas às variáveis quantitativas, caso o usuário responda que a variável é qualitativa, uma nova tela (PreProcesso 1.1) é apresentada informando isto ao usuário. Neste momento ele pode sair do Módulo ou repensar em uma variável de modo que esta seja quantitativa.

A Fig. 4.9 mostra o diagrama até o ponto onde é informado que a variável é qualitativa. O usuário decide recomeçar, retornando a tela PreProcesso1. A seção seguinte é referente a segunda questão.

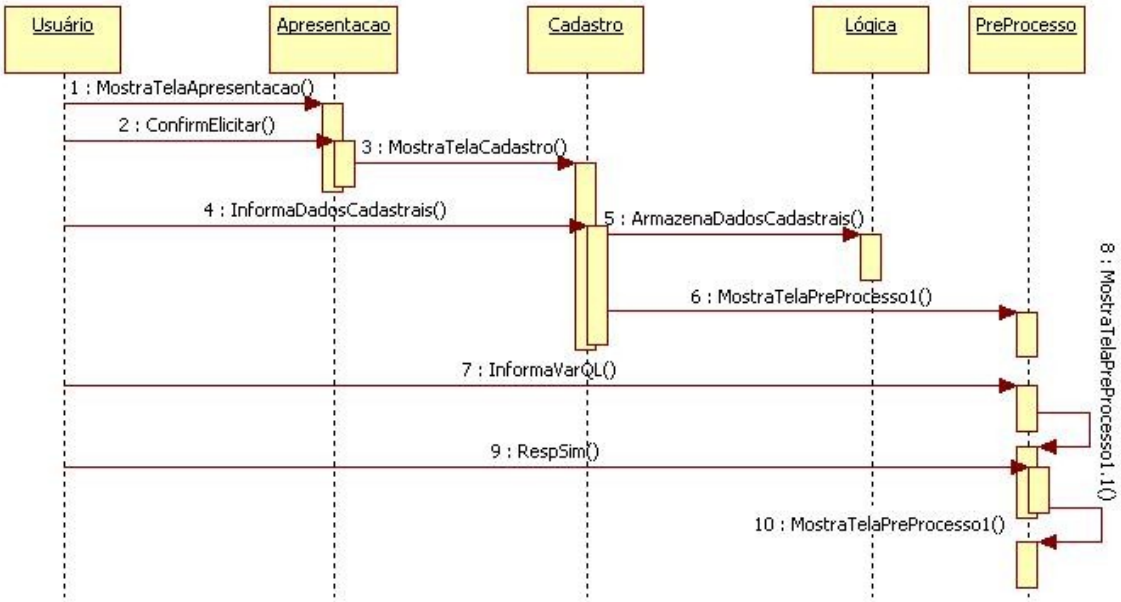

Figura 4.9 – Primeira parte do diagrama de seqüência do fluxo alternativo 6.

4.1.8 – Fluxo Alternativo 7

Caso o usuário responda que a variável é quantitativa, há uma segunda pergunta sobre a classificação da variável feita ao usuário. Esta nova pergunta é se a variável é discreta ou contínua.

O Módulo de Elicitação é apropriado apenas para as variáveis quantitativas contínuas. Portanto, assim como na seção anterior, uma tela (PreProcesso 2.1) é apresentada ao usuário o informando a respeito disso.

Se a variável que ele deseja elicitar não for quantitativa contínua, ele tem a opção de sair do sistema ou repensar em uma variável que obedeça as duas prerrogativas anteriores.

A Fig. 4.10 apresenta o diagrama que ilustra até o ponto que é informado pelo usuário que a variável é discreta. Depois ele decide começar novamente o processo, retornando para a tela PreProcesso1.

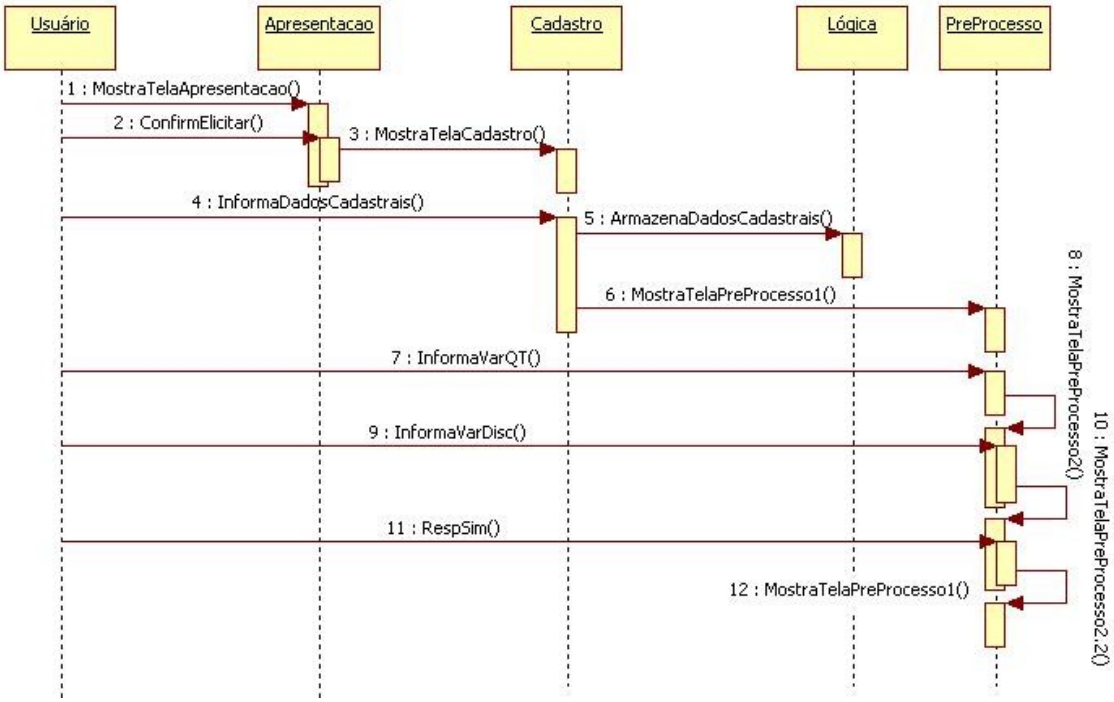

Figura 4.10 – Primeira parte do diagrama de seqüência do fluxo alternativo 7.

4.2 – Interfaces do Módulo de Elicitação

Após a descrição dos vários fluxos que podem existir no Módulo de Elicitação, são apresentadas nesta seção as interfaces, ou telas, do Módulo que correspondem aos objetos apresentados nos diagramas de seqüência.

Cada objeto descrito no diagrama, com exceção do objeto Lógica que é responsável pelo processamento matemático e lógico do Módulo, possui uma tela correspondente.

Como exposto anteriormente, o Módulo de Elicitação começa com uma preparação para o processo de elicitação. Na primeira tela, chamada no diagrama de seqüência de tela Apresentação, há um texto abordando os objetivos deste Módulo, como é o processo de elicitação e quais os resultados esperados.

Esta tela introdutória ao Módulo de Elicitação é vista na Fig. 4.11.

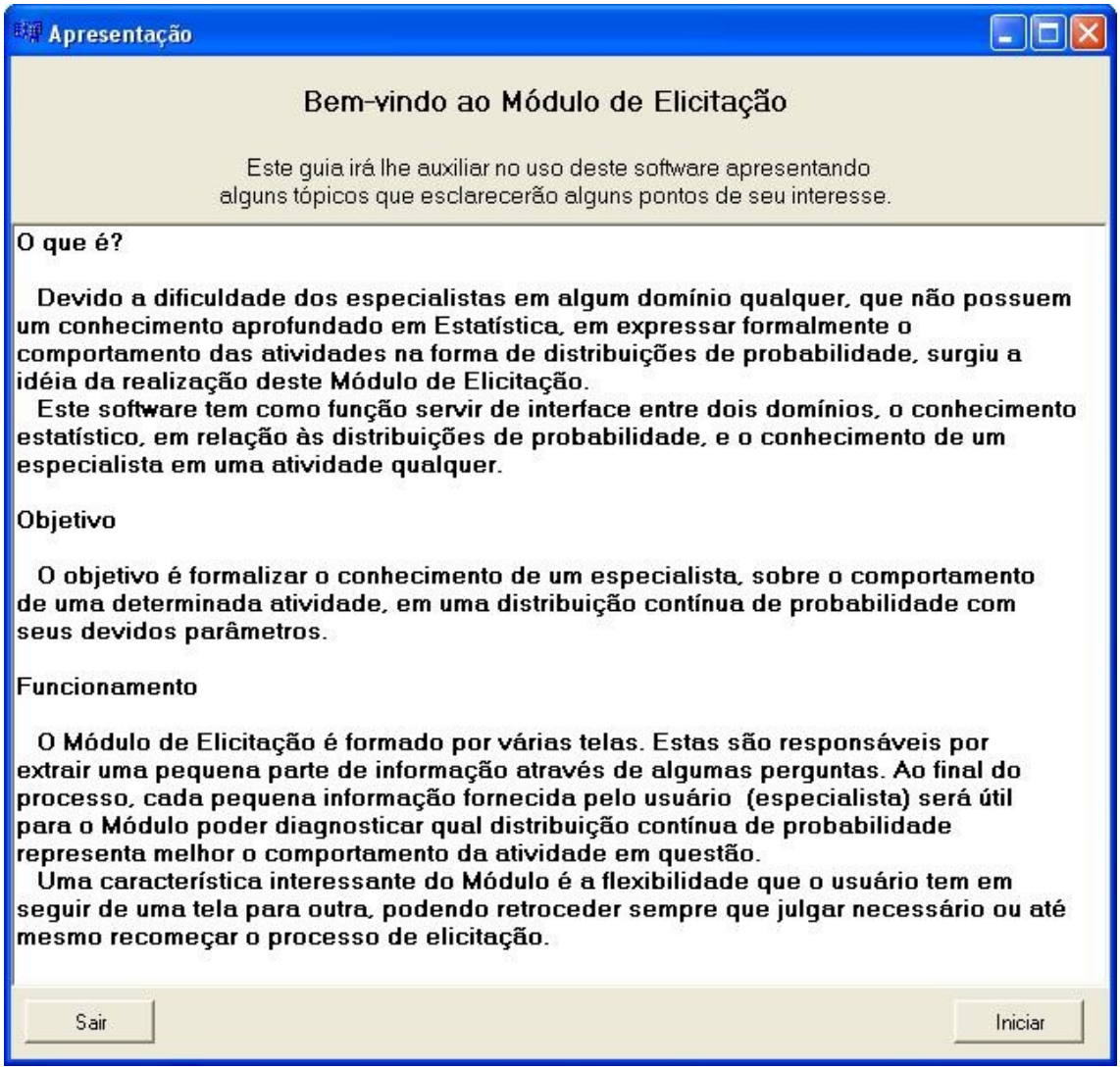

Figura 4.11 – Tela Apresentação.

Em seguida a tela Apresentação, é encontrada a tela Cadastro. Nesta tela é requisitada a identificação do usuário ou especialista, que irá interagir com o Módulo de Elicitação.

A Fig.4.13 representa este formulário inicial no processo de elicitação. As informações foram preenchidas com o intuito de observar como elas são dispostas nas demais telas.

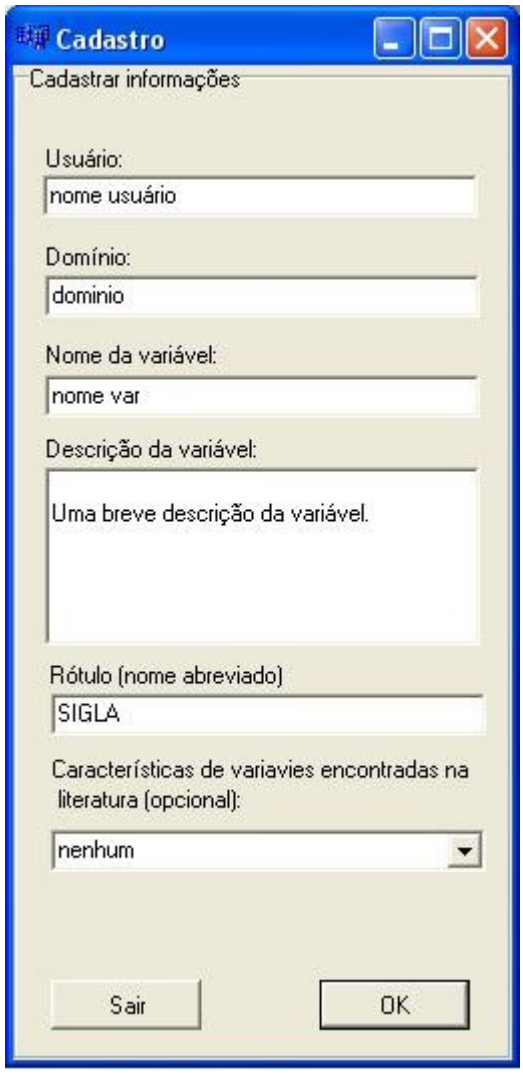

Figura 4.13 – Tela Cadastro.

Além do nome do especialista, outras informações são pedidas, porém nenhuma delas é obrigatória para o seguimento do processo de elicitação.

As demais informações relacionadas à tarefa são: o domínio de aplicação, ou seja, em que contexto esta atividade está inserida.

Uma breve descrição da atividade, que possa esclarecer um pouco sobre a atividade que está sob análise.

Em seguida é requisitado o nome da atividade. Caso seja um nome extenso, há um campo no formulário denominado de rótulo, que serve para abreviar este nome. É pelo nome deste rótulo que a atividade será referenciada no decorrer do processo de elicitação.

A última informação a ser fornecida é em relação à característica da variável. Se esta variável possui relação com alguma característica comumente encontrada na literatura. Pois variáveis com estas mesmas particularidades tendem a apresentar a mesma distribuição de probabilidade.

Na Fig. 4.12 é possível ver as opções de características que são disponibilizadas ao usuário.

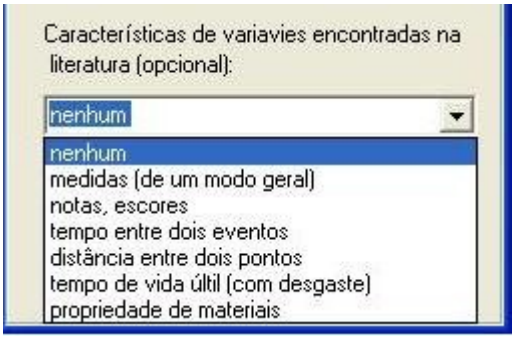

Figura 4.12 – Opções de características encontradas na literatura.

Nesta tela Cadastro, o usuário tem duas opções. Ou segue no processo de elicitação, preenchendo ou não o formulário, ou sai do programa.

Com a inserção das informações, a tela PreProcesso 1 é então a próxima a ser apresentada ao usuário.

Esta tela, assim como a tela PreProcesso 2, que pertencem ao grupo de telas iniciais, é composta por uma pergunta a respeito da variável sob análise com duas possibilidades de resposta, sim ou não. Para auxiliar na resposta, é acomodado ao lado direito da tela, um texto de ajuda que contém a definição e exemplos de aplicação dos termos que estão sendo perguntados.

Como o Módulo de Elicitação se aplica somente as variáveis quantitativas, é necessária a verificação se a variável sob análise atende a este requisito. A Fig. 4.14 é uma ilustração da tela que faz esta verificação.

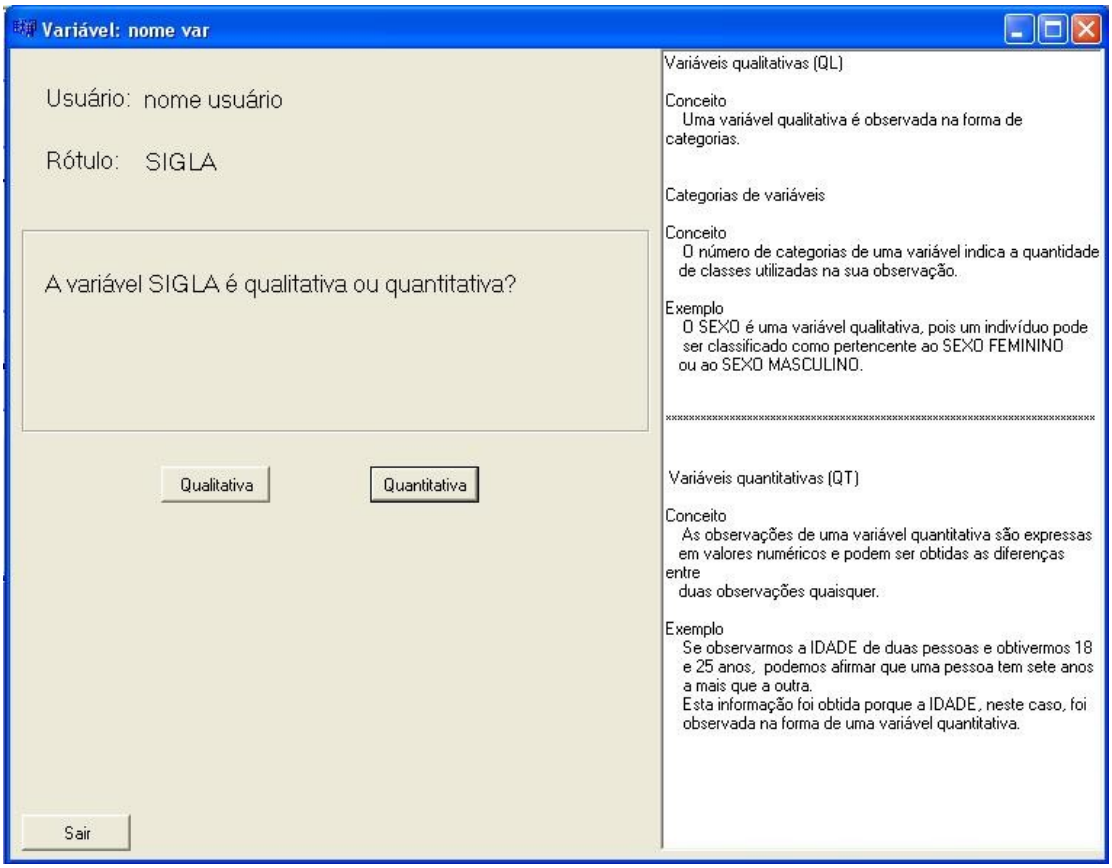

Figura 4.14 – Tela PreProcesso 1.

Caso a variável seja qualitativa, a tela PreProcesso 1.1 é apresentada ao usuário. Esta tela é vista na Fig. 4.15. Se o usuário não desejar recomeçar o processo com uma variável quantitativa, o Módulo é encerrado.

Se a variável for quantitativa, a próxima tela, de acordo com o diagrama de seqüência, é a tela PreProcesso 2 (Fig. 4.16).

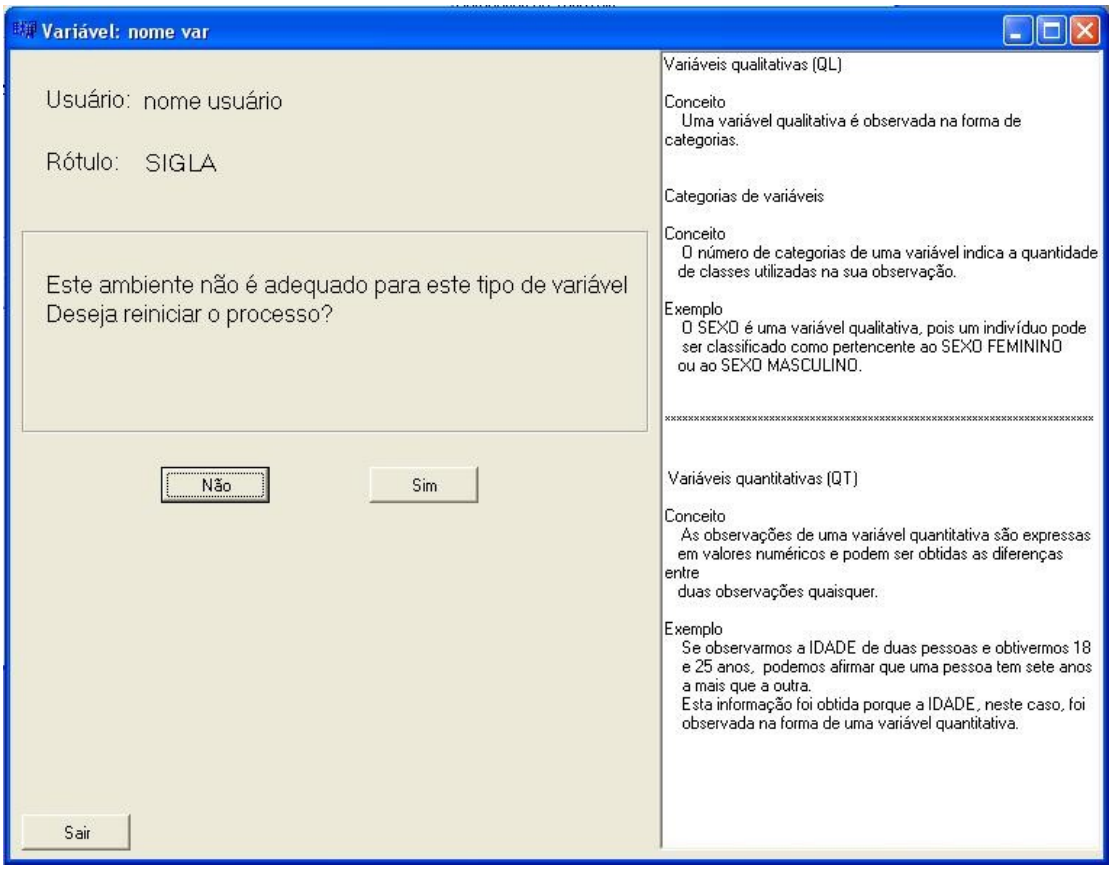

Figura 4.15 – Tela PreProcesso 1.1.

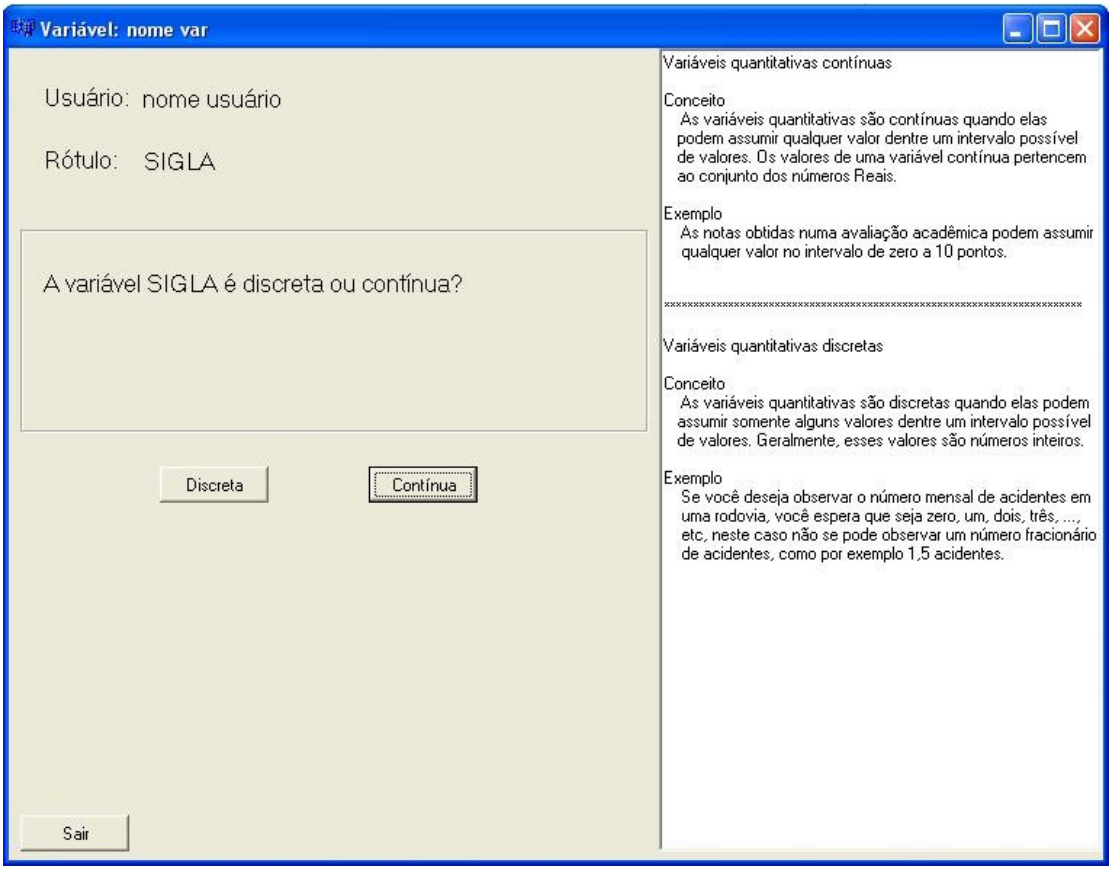

Figura 4.16 – Tela PreProcesso 2.

Na tela PreProcesso 2, uma outra restrição é apresentada. Mesmo sendo uma variável quantitativa, o Módulo é específico para as quantitativas contínuas. Portanto, se a variável que o usuário está analisando não satisfizer essas condições, o Módulo não funciona adequadamente.

Semelhantemente a tela PreProcesso 1.1 apresentada na Fig. 4.15, a tela PreProcesso 2.1 (Fig. 4.17) também apresenta um questionamento sobre o reinício do processo, ainda disponibilizando de um texto auxiliar com os conceitos em questão.

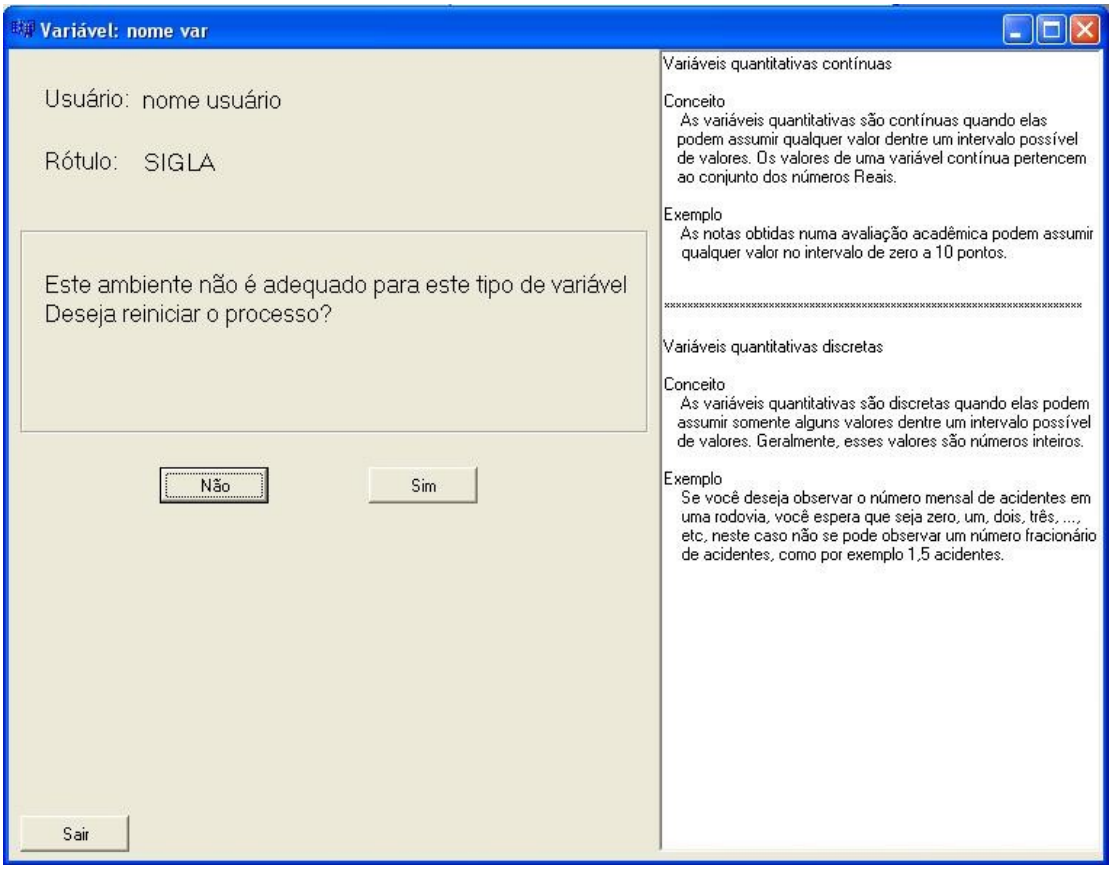

Figura 4.17 – Tela PreProcesso 2.1.

A partir deste momento, o Módulo é apropriado para o uso da variável em análise. Então o processo de elicitação da distribuição contínua de probabilidade se apresenta de uma forma mais clara.

Conforme apresentado nesse estudo, a etapa de extração da informação contida no processo de elicitação foi divida em duas etapas. A primeira referente à descoberta da forma da distribuição de probabilidade. E a segunda diz respeito ao cálculo dos parâmetros desta distribuição.

Conseqüentemente, as telas também foram dispostas nesta ordem. A tela desta primeira etapa está relacionada às perguntas dispostas na forma de uma árvore de decisão, a qual já foi mostrada neste estudo.

A tela Perguntas dispõe as perguntas uma de cada vez, dependendo da pergunta que o objeto Lógica forneceu.

Nesta tela existem quatro possibilidades de resposta para o usuário. Este pode responder afirmativamente para a pergunta e seguir adiante clicando no botão “sim”. Pode discordar e responder negativamente, clicando no botão “não”. Pode voltar à pergunta anterior, clicando no botão “anterior”, caso ache que tenha errado na resposta de uma pergunta, podendo desta forma traçar um novo caminho. Ou por último, lhe é dada a opção de a qualquer momento encerrar o processo de elicitação, clicando no botão “sair”.

A Fig. 4.18, que contém a pergunta inicial da árvore de decisão, serve como exemplo deste tipo de tela. As demais telas, correspondentes às outras perguntas, seguem o mesmo padrão, mudando apenas o conteúdo do texto.

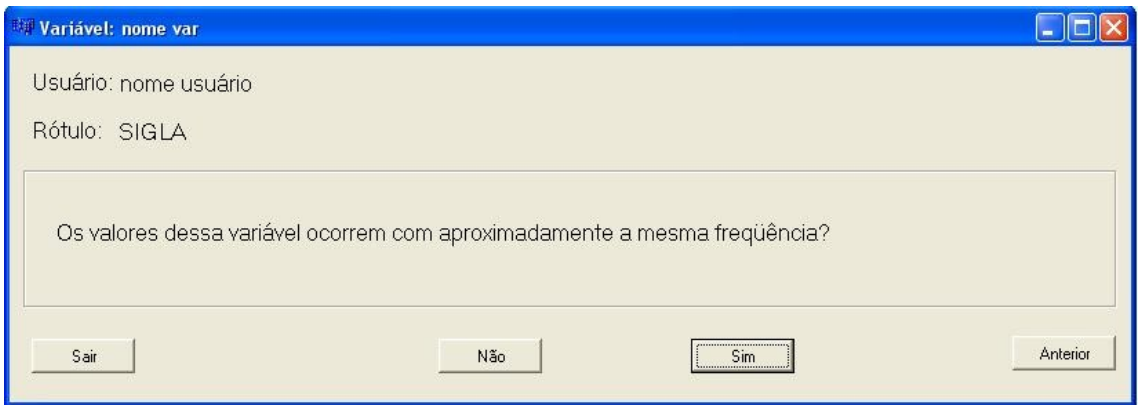

Figura 4.18 – Tela Perguntas: PerguntaIni.

Seguindo o fluxo de perguntas da árvore de decisão, quando a última pergunta é respondida, significa que uma das distribuições de probabilidade foi apontada pelas características fornecidas pelo usuário.

A tela apresentada na Fig. 4.19, chamada no diagrama de seqüência de FigDist, ilustra como esta informação é passada ao usuário quando ou usuário não informou nenhuma característica na tela Cadastro, ou quando a característica da variável fornecida pelo usuário possui a mesma distribuição de probabilidade que a variável elicitada.

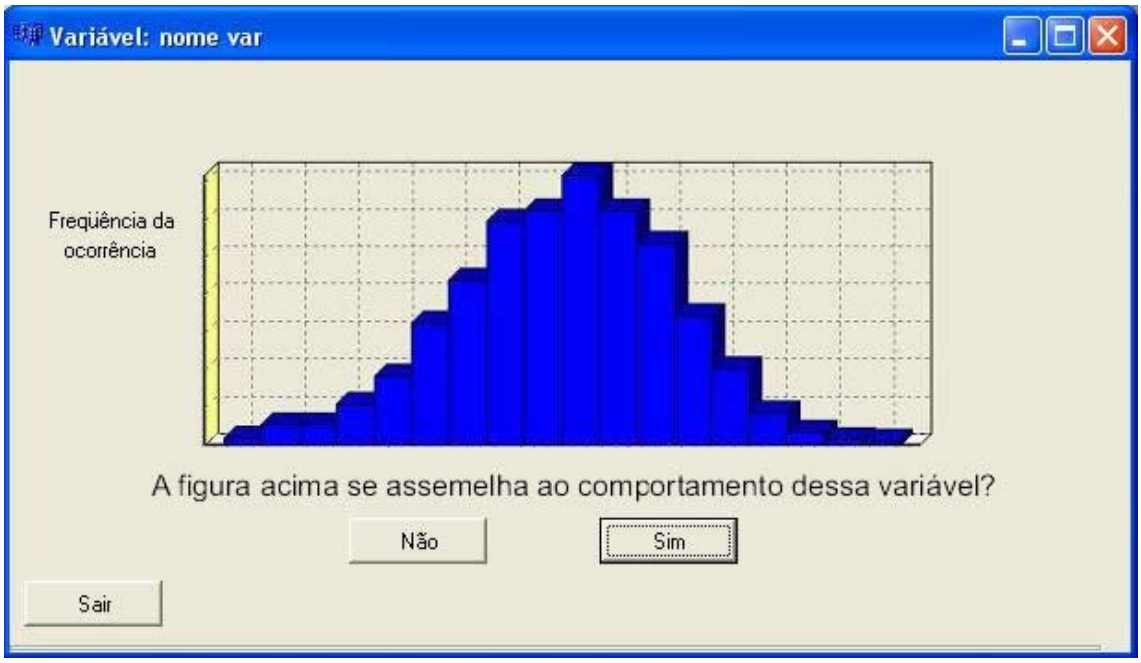

Figura 4.19 – Tela FigDist: sem observação.

Um gráfico em forma de histograma correspondente ao modelo de distribuição padrão é apresentado como um auxílio na confirmação da suspeita de qual distribuição realmente melhor representa a atividade em análise.

Em situações diferentes a essa, onde o usuário tenha informado uma característica na tela Cadastro e a distribuição de probabilidade associada a essa característica não for igual a distribuição elicitada, a tela FigDist conterá algumas informações extras.

Adicionalmente as informações apresentadas na tela da Fig. 4.19, é informado ao usuário que na literatura atual, variáveis com essa característica seguem outro modelo de distribuição.

Para exemplificar, supondo que foi informado na tela Cadastro que a variável analisada tem a característica “tempo entre dois eventos’. A tela da FigDist teria o aspecto visto na Fig. 4.20.

O exemplo mencionado se refere a uma distribuição Normal. Então, de acordo com a distribuição sugerida, diferentes perguntas a respeito dos parâmetros são mostradas na tela Parâmetros para o usuário. Variando de acordo com os valores necessários para o cálculo dos parâmetros reais da distribuição.

Seguindo com o exemplo da distribuição Normal, a tela para aquisição dos valores necessários para o cálculo de μ e σ é apontada na Fig. 4.21.

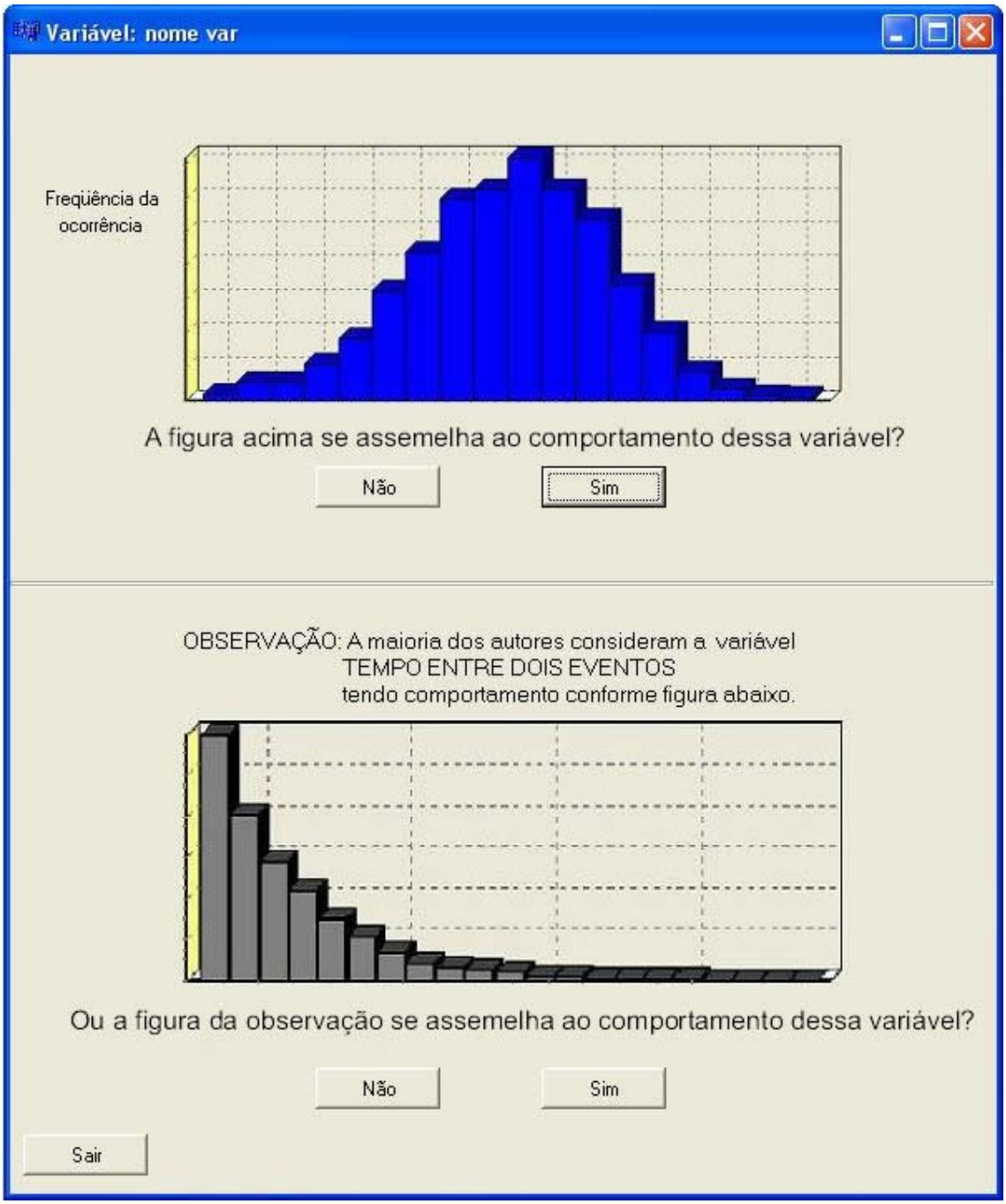

Figura 4.20 – Tela FigDist: com observação.

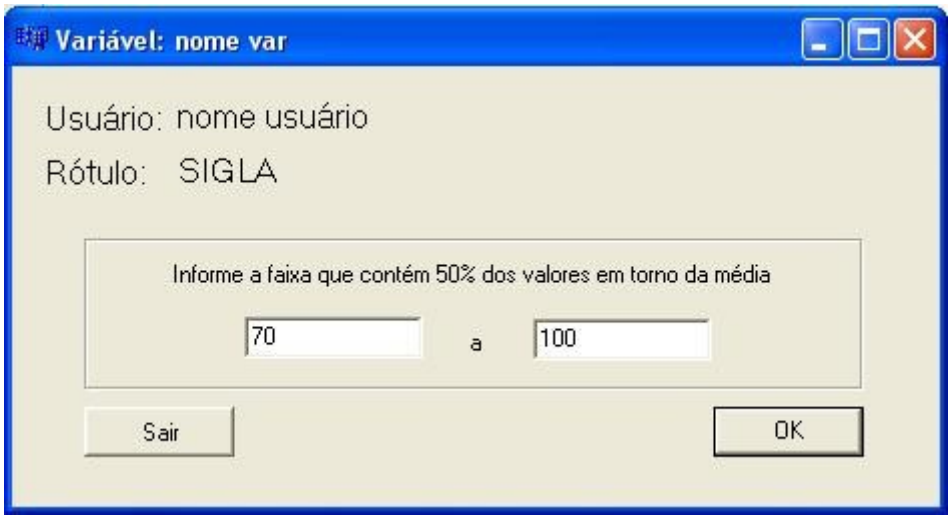

Figura 4.21 – Tela Parâmetros 1.

Depois de informados os valores necessários uma nova tela, chamada de Parâmetros 2 (Fig. 4.22), é apresentada contendo uma simulação da distribuição encontrada utilizando os parâmetros calculados. O Método de Monte Carlo, visto na seção 2.9.1., foi utilizado para esta simulação.

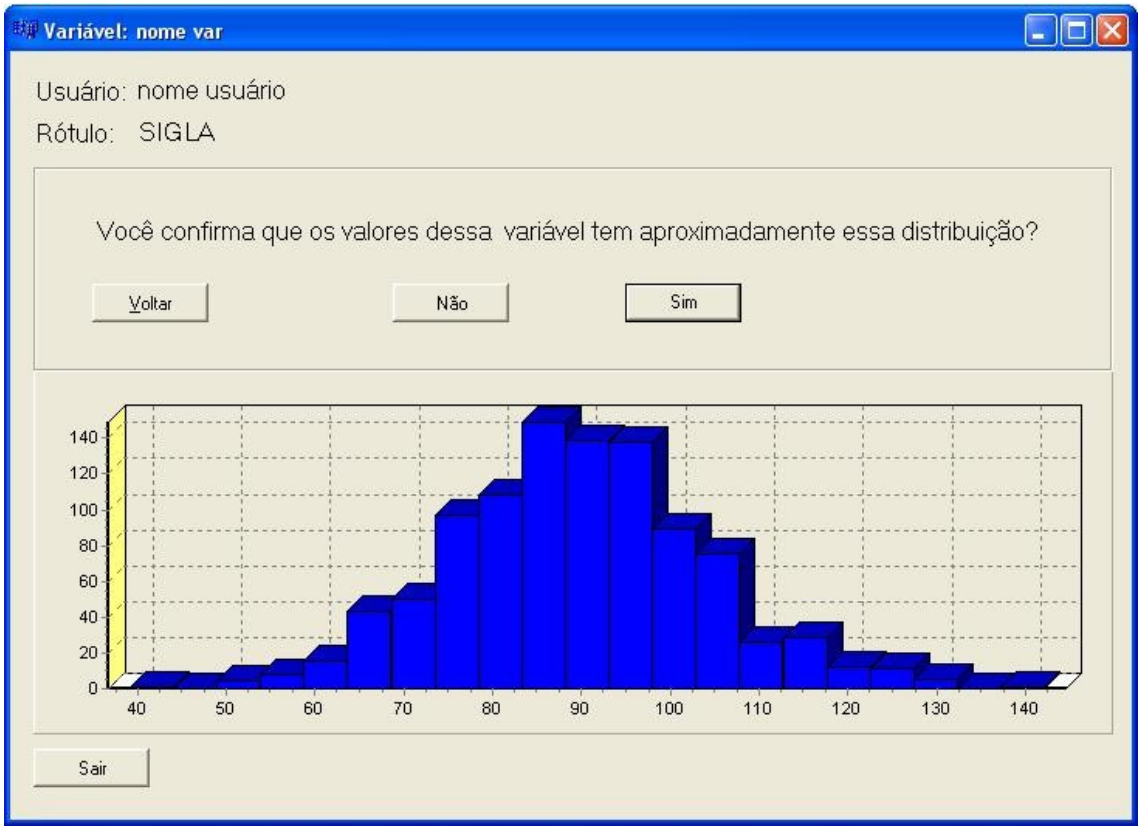

Figura 4.22 – Tela Parâmetros 2.

O usuário tem então, na tela Parâmetros 2 a opção de confirmar se a forma e os parâmetros da distribuição estão de acordo com suas estimativas, utilizando o histograma com a escala de valores reais da variável como referência. Ou ele pode retornar para a tela anterior, Parâmetros 1, e redefinir os valores requisitados. Ou pode negar que o gráfico que ele está vendo representa o comportamento da variável.

Neste último caso, a tela NaoConfirma (Fig. 4.23) é acionada para se saber que ação o usuário deseja efetuar.

Dentre as ações permitidas nesta tela estão: cancelar a última ação, ou seja, voltar para a tela Parâmetros 2 com a possibilidade de seguir adiante ou efetuar qualquer outra opção válida nesta tela.

Pode também recomeçar o processo de elicitação, voltando para a tela Cadastro. E a última alternativa é sair do Módulo.

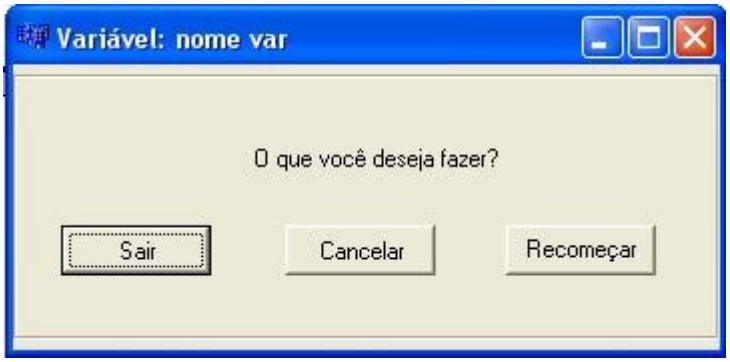

Figura 4.23 – Tela NaoConfirma.

Confirmando que o gráfico da tela representada pela Fig 4.22 está adequado à realidade da variável, então a tela Conclusão, que é a última tela do Módulo de Elicitação, é apresentada. Ela pode ser observada na Fig. 4.24.

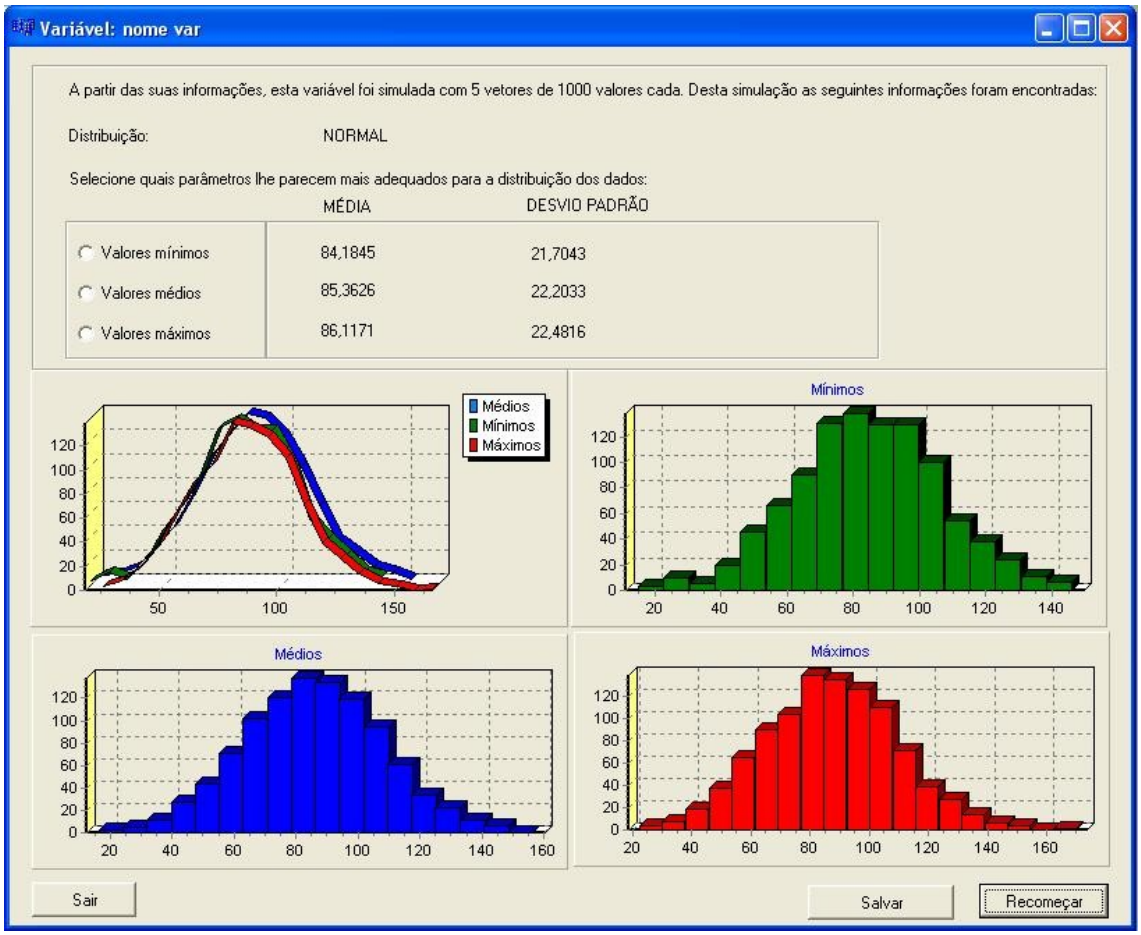

Figura 4.24 – Tela Conclusão.

Nesta última tela estão contidas as informações finais do Módulo de Elicitação, que é constituída do modelo teórico de distribuição juntamente com seus parâmetros.

Os parâmetros recebem maior atenção neste momento, pois ao usuário é dada a possibilidade de três opções, são elas: escolher os valores mínimos, médios ou máximos.

A cada uma dessas opções é apresentado um gráfico correspondendo a visão que o usuário tem da variável. Os valores mínimos são formados pelos menores valores de cada posição dos cinco vetores de dados simulados. Já os valores médios são compostos pela média dos valores de cada posição dos cinco vetores. E finalmente, os valores máximos são constituídos dos valores máximos de cada posição do vetor.

Estas três visões estão relacionadas com o perfil tanto do usuário quanto o perfil da variável. Permitindo que o usuário escolha dentre as três possibilidades qual reflete o comportamento ou previsão da variável. Se é realista, pessimista ou otimista.

Adicionalmente, existe um quarto gráfico que ilustra as três visões concomitantemente, permitindo assim que se possa fazer um paralelo entre todas as possíveis abordagens.

O resultado final é um arquivo texto (Fig. 4.25) contendo as características informadas na tela Cadastro juntamente com o modelo de distribuição de probabilidade elicitado e os parâmetros da visão adequada, que foi escolhida pelo usuário.

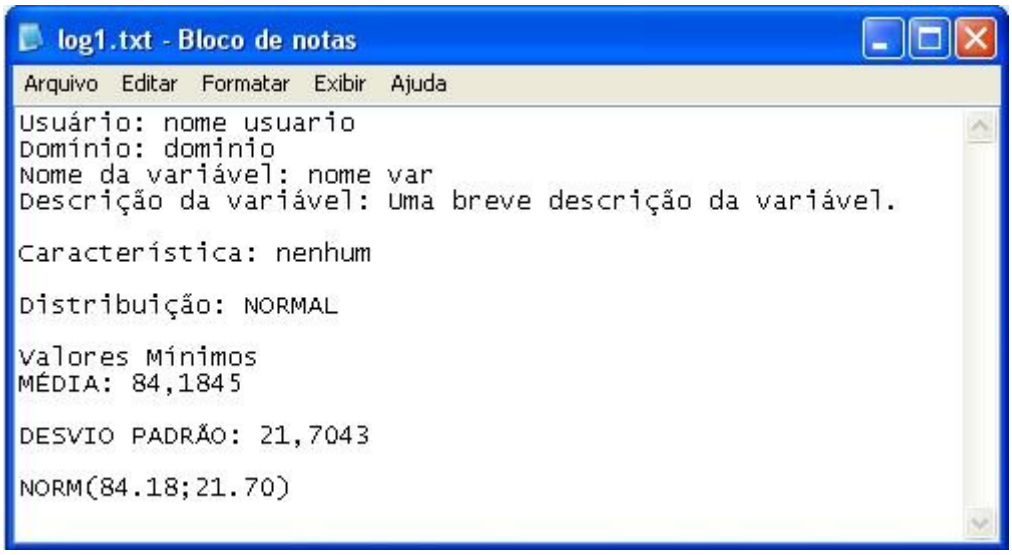

Figura 4.25 – Arquivo texto das informações elicitadas.

Capítulo 5

Resultados

Neste capítulo são apresentadas as abordagens que foram utilizadas no processo de validação tanto da descoberta da forma da distribuição como do cálculo dos parâmetros.

Após a descrição da abordagem adotada, são apresentados os resultados deste processo.

5.1 – Validação do processo de descoberta da forma da distribuição

A etapa de validação do processo de descoberta da forma da distribuição, apresentada neste estudo, consiste em verificar se o processo de elicitação da forma da distribuição implementado no Módulo de Elicitação consegue abstrair as informações necessárias para a descoberta da distribuição.

Para isto, verifica se as perguntas estão formuladas de forma clara, de forma que o especialista compreenda que tipos de características estão sendo perguntados.

O processo de validação foi estruturado da seguinte maneira:

- Escolha da variável que o especialista desejava elicitar.
- As figuras das distribuições foram dispostas em um papel e o especialista identificou qual delas representava o comportamento da variável.
- Utilização do Módulo de Elicitação.
- Comparação entre o resultado do Módulo e a figura indicada no passo 2.

Este processo foi submetido três vezes para cada um dos cinco especialistas nos mais diversos domínios, desde a área médica até a área de planejamento e distribuição de energia elétrica.

Nas situações onde a distribuição de probabilidade escolhida no início do processo eram as distribuições Triangular ou Uniforme, os especialistas rapidamente chegaram a estas distribuições na conclusão do Módulo.

Quando a distribuição a ser encontrada era a distribuição Normal, nem sempre ela foi a indicada pelo Módulo de Elicitação. Em alguns casos a distribuição Triangular também foi encontrada.

As distribuições Lognormal e Exponencial também foram bastante confundidas. Porém, a maior confusão estava na identificação do especialista em qual das duas realmente representava a variável analisada.

Após a elicitação, mesmo o Módulo tendo apontado para uma distribuição diferente da sugerida em princípio pelo especialista, este confirmou que a distribuição apontada pelo Módulo refletia melhor o comportamento da variável do que a indicação da distribuição anterior.

Poucos casos foram atribuídos tendo uma distribuição Weibull, no entanto as vezes que esta foi indicada, o Módulo de Elicitação teve êxito ao encontrá-la.

A distribuição Beta, apesar de não ter sido implementada uma solução para o cálculo dos seus parâmetros, estava presente na validação da descoberta da forma. Algumas variáveis escolhidas pelos especialistas possuíam sua característica de distribuição e foram permitidas participar do processo de validação. Todas as vezes que ela foi escolhida o Módulo conseguiu identificá-la.

Em alguns casos, os especialistas que utilizaram o Módulo apresentaram dificuldades durante o processo e propuseram mudanças para melhorá-lo. Estas mudanças foram acatadas e implementadas, constando agora na versão final do Módulo.

Na tela FigDist do Módulo de Elicitação que apresenta um gráfico para o especialista confirmar se a forma da distribuição é realmente aquela, duas situações de apresentação eram possíveis. Ou apenas um gráfico era visível, ou dois gráficos eram mostrados. Um correspondendo à distribuição elicitada e um relacionado com a característica da variável.

Nessas situações, todas as duas alternativas foram escolhidas pelos especialistas. Em certos casos, mesmo com a indicação de que comumente na literatura aquele tipo de variável era representada por uma distribuição diferente daquela elicitada, o especialista confirmou que a distribuição elicitada era a melhor representação do comportamento da variável.

Em outros casos, a indicação da observação a respeito da variável contribuiu para que o especialista repensasse na forma da distribuição e reconsiderasse ou escolher a distribuição da observação ou recomeçar o processo.

5.2 – Validação do cálculo dos parâmetros da distribuição

A validação do cálculo dos parâmetros consiste em analisar se a partir dos valores informados pelo especialista, os parâmetros encontrados em conjunto com o modelo teórico possam representar com fidelidade o comportamento do modelo real.

Para o processo de validação foram utilizadas duas abordagens. A primeira utiliza o conhecimento e experiência dos estatísticos para avaliarem o cálculo proposto. A segunda abordagem se vale de técnicas de simulação para que a partir dos parâmetros calculados seja possível encontrar os valores informados pelo especialista.

5.2.1 – Validação pelos especialistas

Um grupo de cinco estatísticos foi convidado para participar de uma banca e a eles foi entregue um documento contendo todos os cálculos dos parâmetros.

Após a apresentação do trabalho, estes levaram o documento para uma análise mais detalhada. Em seguida o documento foi entregue com a aprovação dos cálculos e alguns ajustes, os quais foram acatados e já fazem parte da solução apresentada.

5.2.2 – Validação por simulação

Juntamente com o processo de validação descrito na seção anterior, uma outra abordagem foi utilizada. Esta utilizando o Método de Monte Carlo.

Foram mostrados, na seção 4.2, os cálculos realizados para se chegar aos valores dos parâmetros. Sempre partindo de valores mais intuitivos, para um especialista sem muitos conhecimentos estatísticos, para a descoberta dos parâmetros reais da distribuição.

A idéia de validar este processo foi utilizar o caminho oposto. Partindo de uma distribuição gerada, por meio de simulação, com parâmetros conhecidos, para obter os valores que são perguntados ao especialista.

Após o especialista ter informado os valores requisitados pelo módulo de elicitação e o módulo ter calculado os parâmetros, a FGVA da distribuição encontrada irá gerar um vetor de dados utilizando como parâmetros os valores calculados pelo módulo.

Este vetor de dados será então submetido a uma análise matemática e estatística para descobrir qual o valor nesse vetor em relação ao valor que o especialista informou no processo de elicitação.

Caso os valores dessas informações sejam iguais ou próximos, isto significará que o processo de cálculo dos parâmetros realmente obtém os valores corretos.

A Fig. 5.1 ilustra o processo de validação dos parâmetros por meio da simulação

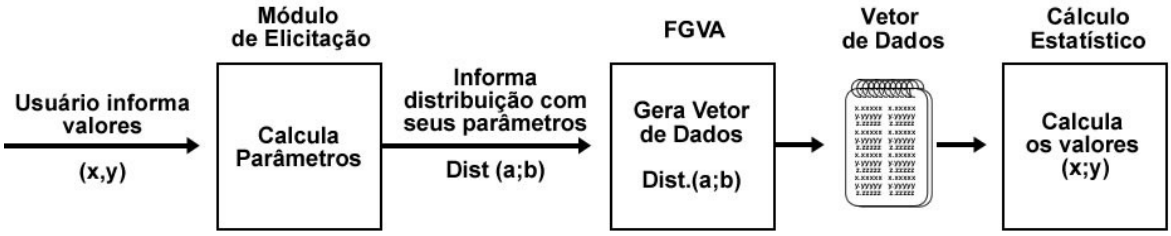

Figura 5.1 – Processo de validação dos parâmetros por meio de simulação.

No caso das distribuições, Uniforme e Triangular, estas não necessitam de tal validação. Isto decorre do fato de que os parâmetros, no caso da distribuição Uniforme, são os mesmos valores que são perguntados. Em relação à distribuição Triangular, apenas a moda é calculada, entretanto é obtida facilmente, sem nenhum cálculo de maior complexidade.

Já no que diz respeito às demais distribuições, estas sim merecem esta verificação. Outra importância deste teste é em relação à compreensão da distância máxima e mínima que pode existir entre os parâmetros informados pelo especialista.

Dependendo da distância, pode ser necessário alertá-lo das conseqüentes mudanças que a distribuição pode apresentar. Podendo-se também verificar os limites em que a solução proposta é válida.

Para a geração do vetor de dados pelo Método Monte Carlo e para a descoberta dos valores obtidos pela FGVA, foi utilizado o software EPRisk 4, desenvolvido pela equipe do PerfomanceLab do Departamento de Informática e Estatística da Universidade Federal de Santa Catarina em parceria com a empresa estatal Petrobrás S.A.

5.2.2.1 – Distribuição Normal

A primeira distribuição a ser apresentada utilizando o processo de validação descrito será a distribuição Normal.

No módulo de elicitação, a distribuição Normal necessita que o especialista informe a faixa de valores que compreende os 50% dos dados em torno da média. Portanto é fornecido um valor referente ao limite inferior e um valor ao limite superior.

Os parâmetros reais da distribuição, como apontado na seção 3.2.3, são: a *média* (μ) e o *desvio padrão* (σ).

No teste realizado, os limites, inferior e superior, receberam diversos valores. Começando bem próximos e se distanciando até chegarem a um intervalo muito grande.

Uma vez informados os limites mencionados, o cálculo proposto neste estudo para a descoberta dos parâmetros da distribuição Normal é executado a fim de se encontrar os valores de μ e σ.

A FGVA da distribuição Normal, utilizando os parâmetros calculados, gera cinco vetores de dados com 1000 valores cada. Estes vetores são ordenados e em seguida é calculada a média do valor encontrado na posição 250, que corresponde ao limite inferior, e na posição 750, que se refere ao limite superior. A finalidade de se calcular a média de cinco vetores é diminuir pequenos erros, que são valores um pouco fora da faixa esperada, na geração dos valores. Utilizando a média desses valores, é esperado que estes valores se comportem adequadamente.

O objetivo é avaliar se estes limites estão de acordo com os limites fornecidos pelo especialista inicialmente. A Tabela 5.1 mostra os resultados obtidos com estas simulações.

A primeira coluna da Tabela 5.1 é referente aos valores dos limites, inferior e superior, que foram fornecidos pelo especialista. A segunda coluna são os valores dos parâmetros da distribuição calculados pelo módulo de elicitação. E na terceira coluna constam os valores correspondente aos fornecidos pelo especialista, obtidos a partir do vetor de dados gerado pela FGVA da distribuição Normal com os parâmetros de μ e σ que o especialista informou.

Tabela 5.1 – Variação da distância entre o limite inferior e superior da distribuição Normal

Limites informados		**Parâmetros calculados**		**Limites encontrados na simulação** (Média de cinco vetores gerados)	
Inferior	Superior	μ	σ	Inferior	Superior
50	55	52,56	3,7	50,02	55,04
40	60	50	14,83	40,1	60,2
25	75	50	37,07	25,26	75,5
36	120	79,01	62,17	36,44	120,64
100	200	150	74,14	100,52	201
20	150	86,57	96,21	20,67	151
15	250	135,34	173,93	16,2	251,8
133	490	315,5	264,68	134,85	493,57
80	600	346,28	384,86	82,7	604
50	1000	536,48	703,1	54,93	1007,29

Observando a Tabela 5.1, pode-se notar que os limites encontrados são bem próximos dos fornecidos pelo especialista. Demonstrando que o cálculo de μ e σ está sendo realizado com êxito no Módulo de Elicitação.

Porém com um intervalo muito grande, o que não é muito comum em atividades reais, os limites calculados começam a apresentar um pouco de distorção em relação ao informado inicialmente.

Como se trata de intervalos muito grandes, mesmo esta pequena distorção não parece trazer qualquer tipo de prejuízo ao objetivo final, que é o cálculo dos parâmetros.

5.2.2.2 – Distribuição Lognormal

A próxima distribuição é a distribuição Lognormal e seus parâmetros são: *média* e *desvio padrão*. Para calculá-los é pedido ao especialista o valor mínimo e a moda, como apresentado na seção 3.2.4.

Para esta simulação, o valor mínimo foi fixado em quatro e a moda foi variando tendo como referência o valor do mínimo. Esta variação mostrou qual a relação da moda com o valor mínimo. Em seguida outros valores de mínimo e moda foram simulados.

A FGVA da distribuição Lognormal, assim como a distribuição Normal e as demais, gera cinco vetores de dados com 1000 valores cada vetor. Depois dos valores estarem ordenados, a média do primeiro elemento dos cinco vetores é calculada. Para o

cálculo da *moda* foi verificado no vetor qual classe possui maior concentração de valores.

A Tabela 5.2 apresenta o comportamento dos parâmetros e a comparação entre eles. Na primeira coluna estão os valores, mínimo e moda, informados pelo especialista. Na segunda coluna estão os parâmetros calculados pelo módulo elicitação. Na terceira coluna estão os valores de mínimo e moda, encontrados no vetor gerado pela FGVA.

Tabela 5.2 - Variação da distância entre o valor mínimo e o valor da medida moda.

Mínimo e moda informados		**Parâmetros calculados**		**Mínimo e moda encontrados na simulação** (Média de cinco vetores gerados)	
Mínimo	Moda	Média	Desvio	Mínimo	Moda
4 (9x)	36	5,02	1,27	4,74	88,7
4 (8x)	32	4,65	1,09	4,62	55,6
4 (7x)	28	4,23	0,95	4,54	40,5
4 (6x)	24	3,85	0,82	4,46	31,5
4 (5x)	20	3,49	0,7	4,39	22,4
4 (4x)	16	3,1	0,57	4,31	16,3
4 (3x)	12	2,67	0,43	4,23	13,2
4 (2x)	8	2,14	0,25	4,13	10,1
89	108	4,69	0,06	89,8	107,8
70	98	4,6	0,12	71,06	95,1
50	85	4,48	0,19	51,27	81,3
40	80	4,44	0,25	41,36	83,4
33	100	4,79	0,43	34,95	95
60	235	5,77	0,56	64,65	232,6
18	84	4,86	0,65	19,65	97,6
15	100	5,42	0,9	16,9	136,9
7	59	5,41	1,15	8,16	64

De acordo com os valores apresentados na Tabela 5.2 é possível perceber que quando o valor da moda for maior que cinco vezes o valor do mínimo, a moda encontrada no vetor gerado não confere com a moda informada pelo o especialista.

Quando o valor da moda equivale a dez vezes o valor do mínimo, a subtração dentro da raiz quadrada da solução proposta, gera um valor negativo. Portanto não é possível realizar tal operação.

Nos demais casos há uma aproximação satisfatória dos valores. Pequenas variações podem ser justificadas ou pelo processo de simulação, apesar se estar utilizando a média de cinco vetores. Ou ao fato do cálculo da moda em distribuições contínuas ser baseado em valores médios das classes, perdendo um pouco da precisão dos valores.

5.2.2.3 – Distribuição Exponencial

A distribuição Exponencial apresenta como parâmetro o valor λ. Como este parâmetro é o inverso da *média*, para uma maior simplicidade, os cálculos foram baseados na *média* e não no λ.

Para o cálculo da média são necessários o valor mínimo e a mediana, como apresentado na seção 3.2.5. Desta forma, as simulações realizadas verificam o comportamento da distribuição quando a mediana se distancia do valor mínimo.

Após a geração dos vetores de dados, é encontrado no vetor ordenado, referente à média de todos os vetores, a posição 0, que indica o menor elemento. E a posição 500, que está no meio do vetor, contendo metade dos valores inferiores a ele e a outra metade valores superiores.

A Tabela 5.3 apresenta em sua primeira coluna os valores, mínimo e mediana que o especialista forneceu. A segunda coluna contém o parâmetro que foi calculado baseado nas informações do especialista. E na terceira coluna estão os valores encontrados no vetor de dados gerados pela FGVA da distribuição Lognormal.

Tabela 5.3 – Variação da distância entre o valor mínimo e o valor da mediana.

Mínimo e mediana informados		**Parâmetro calculado**	**Mínimo e mediana, encontrados na simulação** (Média de cinco vetores gerados)	
Mínimo	Mediana	Média	Mínimo	Mediana
4 (250x)	1000	1436,9	5,71	1020
4 (100x)	400	571,3	4,68	407,9
4 (50x)	200	282,77	4,34	203,9
4 (30x)	120	167,35	4,2	122,32
4 (25x)	100	138,5	4,16	102
4 (20x)	80	109,64	4,13	81,52
4 (15x)	60	80,79	4,1	61,12
4 (10x)	40	51,94	4,06	40,72
4 (9x)	36	46,17	4,05	36,64
4 (8x)	32	40,39	4,05	32,56
4 (7x)	28	34,62	4,04	28,48
4 (6x)	24	28,85	4,03	24,4
4 (5x)	20	23,08	4,03	20,32
4 (4x)	16	17,31	4,02	16,24
4 (3x)	12	11,54	4,01	12,16
4 (2x)	8	5,77	4	8,08

Os valores encontrados na Tabela 5.3 mostram que existe uma tolerância, ao distanciamento entre o valor mínimo e a mediana, muito grande. Até a proporção de dez vezes sendo a mediana maior que o valor mínimo os valores encontrados são precisos.

Após este valor, algumas distorções começam a aparecer, porém ainda mantendo um nível aceitável até se chegar ao valor da mediana ser cem vezes maior que o valor mínimo.

Com uma proporção maior que cem vezes, os valores começam a apresentar diferenças consideráveis entre os valores informados e os encontrados na simulação.

5.2.2.4 – Distribuição Weibull

Conforme apresentado na seção 3.2.6, a distribuição Weibull requer como parâmetros os valores de α e β. E para calculá-los, a solução proposta se vale dos valores máximo e moda.

Portanto, a simulação a seguir identificará para que intervalo entre o valor máximo e o valor da moda, a solução para o cálculo dos parâmetros é válida.

Após o especialista ter informado o valor máximo e a moda, o módulo elicitação calculará α e β. Com esses parâmetros, a FGVA da distribuição Weibull irá gerar o vetor de dados, no qual o valor máximo, localizado na posição 1000 do vetor e a moda serão identificados.

A Tabela 5.4 apresenta os resultados da simulação, constando na primeira coluna os valores, máximo e moda, informados pelo especialista. Na segunda coluna estão os parâmetros calculados. A terceira coluna apresenta os valores equivalentes aos informados, mas que foram localizados no vetor gerado pela FGVA da distribuição Weibull.

O intervalo de aplicação da solução proposta é enquanto o valor da moda for maior que a metade do valor máximo. Pode-se notar na Tabela 5.4 que na situação onde a moda é igual a metade do máximo, o valor de α é menor que 4.

Vale lembrar que a distribuição Weibull que interessa a este estudo é assimétrica a esquerda, e para isto o valor de α deve ser maior que 4.

Tabela 5.4 – Variação da distância entre o valor moda e o valor máximo.

Moda e máximo informados		**Parâmetros calculados**		**Moda e máximo encontrados na simulação** (Média de cinco vetores gerados)	
Moda	Máximo	α	β	Moda	Máximo
95	100	43,74	95,05	95,49	99,25
90	100	21,53	90,20	88,95	99,68
85	100	14,11	85,44	82,74	100,37
80	100	10,4	80,78	81,96	98,5
75	100	8,172	76,21	75,42	98,15
70	100	6,68	71,72	70,45	97,75
65	100	5,61	67,31	65,63	97,32
60	100	4,8	62,99	59,4	96,89
55	100	4,173	58,74	56,62	96,42
50	100	3,66	54,54	49,65	95,94

Para os demais valores de moda que estão localizados no intervalo compreendido entre o valor máximo e a sua metade, os valores encontrados, por meio da simulação, no vetor de dados correspondem aos valores informados pelo especialista.

Entretanto à medida que o valor da moda se aproxima da metade do valor máximo, a moda encontrada no vetor simulado permanece com uma boa precisão, já o valor máximo apresenta uma pequena variação. Porém esta variação não chega a prejudicar a solução.

Capítulo 6

Considerações Finais

6.1 Conclusões

Esta dissertação teve seu foco no aprimoramento do processo de descoberta dos modelos teóricos de distribuições contínuas de probabilidade nas situações onde a disponibilidade de dados para a realização dos testes de aderência é comprometida, ou pela insignificante quantidade de dados ou pela sua inexistência.

Neste contexto, foi apresentada uma proposta de elicitação do conhecimento tácito dos especialistas no domínio da variável de interesse.

A proposta em questão é divida em duas partes, uma responsável pela descoberta da forma da distribuição de probabilidade e outra referente ao cálculo dos parâmetros da distribuição elicitada.

Na primeira etapa, referente à descoberta da distribuição de probabilidade que representa o comportamento da variável, algumas técnicas de formulação de questionários foram utilizadas. Entretanto, as perguntas não estavam em um questionário tradicional, foram aplicadas através do software Módulo de Elicitação, que foi desenvolvido com esta finalidade.

O encadeamento das perguntas estava estruturado em forma de uma árvore de decisão, permitindo que ao final do processo, uma distribuição pudesse ser indicada.

A formulação dessas perguntas foi um processo bastante rigoroso e delicado. Várias vezes foi necessário aplicar as perguntas a pessoas que não tinham domínio em Estatística para avaliar se estas estavam construídas de modo que fosse de fácil entendimento para pessoa leiga ou com poucos conhecimentos estatísticos.

O refinamento do processo de elaboração das perguntas foi contínuo, durando até a fase de validação, onde os especialistas ainda propuseram mudanças, as quais sempre foram acatadas.

A etapa seguinte, relacionada ao cálculo dos parâmetros das distribuições, não foi tão subjetiva quanto a etapa anterior. Estatísticos analisaram as soluções descritas e, após as devidas observações, não se opuseram a nenhuma das técnicas aplicadas.

Concomitantemente à análise dos estatísticos, testes realizados por meio de simulação com os parâmetros calculados mostraram que os cálculos encontrados nesse estudo identificavam adequadamente os parâmetros das distribuições a partir dos valores informados pelo especialista. Entretanto, algumas observações foram feitas relacionadas com a limitação do intervalo que existia entre os valores que o especialista informava.

Após estas duas etapas para a elicitação da distribuição de probabilidade devidamente acompanhada dos valores de seus parâmetros, a preocupação estava em deixar o processo o mais interativo possível, permitindo que o especialista do Módulo de Elicitação pudesse se movimentar entre as telas, as quais correspondem às etapas, do modo que quisesse ou precisasse para conseguir chegar a uma conclusão.

Essa diversificação de fluxo no Módulo foi abordada nesse estudo, deixando claro os possíveis caminhos que poderiam ser percorridos, sempre respeitando a ordem das etapas. Primeiro a descoberta da distribuição de probabilidade e depois o cálculo dos seus parâmetros. E sempre permitindo ao especialista retroceder ou recomeçar o processo.

A validação do Módulo de Elicitação, como um todo, foi um processo bastante complexo. Essa complexidade foi ocasionada pela pressão que os especialistas, que estavam interagindo com o Módulo, se faziam. A principal dificuldade era fazer com que os especialistas compreendessem que quem estava sendo analisado e julgado era o Módulo de Elicitação e não eles próprios.

Após esta barreira inicial superada, que acontecia a partir da segunda elicitação, eles começavam a ser mostrar familiarizados com o processo e seguiam com tranqüilidade até a etapa final do Módulo.

Para cada um dos cinco especialistas que utilizaram o Módulo, três elicitações foram realizadas. Em alguns casos com variáveis diferentes e em outros casos, os especialistas decidiam recomeçar o processo com a mesma variável, uma vez que estes estavam insatisfeitos com o resultado encontrado.

Mas invariavelmente, a segunda elicitação era bem mais simples e ocasionava em um resultado que agradava o especialista. O que comprovou que é necessário a realização de uma rodada de calibração para o sucesso do processo.

No final da interação do especialista com o Módulo de Elicitação, todos os especialistas se mostraram satisfeitos com os resultados de um modo geral e já estavam manuseando o software com propriedade, o que atesta a facilidade no seu uso.

Desta forma é possível concluir que o Módulo de Elicitação obteve êxito na difícil tarefa de extrair o conhecimento tácito dos especialistas e poder formalizar esse conhecimento na forma de distribuições de probabilidade, juntamente com seus parâmetros.

6.1 Trabalhos Futuros

No decorrer de uma pesquisa de dissertação sempre surgem idéias interessantes que norteiam o desenvolvimento do documento final. Entretanto, algumas delas não podem ser aprofundadas e exploradas devido ao curto período de tempo disponível.

Conseqüentemente, estas idéias ficam como uma oportunidade de estudo para aqueles que desejarem dar continuidade à linha de pesquisa desta dissertação.

A principal proposta para trabalhos futuros, dando continuidade a esta pesquisa, é a expansão da solução apresentada para outras distribuições contínuas de probabilidade. Para isto seria necessário redesenhar a árvore de decisão, adicionando outros aspectos da característica das variáveis que não foram abordados neste estudo.

A distribuição Beta foi utilizada apenas na estruturação da árvore de decisão. Embora soluções para o cálculo dos parâmetros desta distribuição já tenham sido propostas por outros autores, no caso deste trabalho a distribuição Beta é utilizada em um único caso especial, dentre todas suas formas. Desta maneira, a solução apresentada por eles não é satisfatória por não restringir a este único caso. E devido à falta de tempo, não foi possível definir um cálculo capaz de encontrar seus parâmetros.

No entanto, a distribuição Beta foi considerada na estruturação da árvore de decisão.

Além dessa nova estruturação da aquisição da forma com a inserção de novas distribuições, seria necessário também estudar a relação dos parâmetros da nova distribuição com pelo menos dois dos valores que já fazem parte da solução, e encontrar um cálculo capaz de fazer esta transformação.

Seria útil também aplicar este procedimento para as distribuições discretas, verificando se a solução encontrada nesta dissertação também é válida nesse escopo.

Com a incorporação de novas distribuições, seria importante o estudo de novas técnicas de elicitação para auxiliar o processo atual. Encorpando o Módulo de Elicitação com ferramentas que possam deixar o especialista ainda mais focado no processo, produzindo resultados tão satisfatórios quantos os resultados encontrados atualmente.

Referências

AYYUB, Bilial M. Elicitation of expert opinions for uncertainty and risks. Florida, USA. CRC Press, 2001.

BABYLON. Disponível em: < http://www.babylon.com>. Acesso em: 20/03/2008.

BARBETTA, Pedro. A. Estatística Aplicada às Ciências Sociais. 6ª edição. Florianópolis: Editora da UFSC, 2006.

BARBETTA, Pedro A.; REIS, Marcelo M. & BORNIA, Antônio C. Estatística para cursos de engenharia e informática. São Paulo. Editora Atlas. 2004.

BITTENCOURT, G. Inteligência Artificial: ferramentas e teorias. 2ª edição. Florianópolis: Editora da UFSC, 2001.

BRATLEY, Paul; BENNETT, Fox L. & SCHRAGE, Linus E. A Guide to Simulation. 2º Edition. USA. Springer-Verlag, 1987.

DANESHKHAH, A. R. Psychological Aspects Influencing Elicitation of Subjective Probability**.** The University Of Sheffield. August, 2004. Disponível em: <www.shef.ac.uk> Acesso em: 15/01/2008.

DANESHKHAH, A. R. Uncertainty in Probabilistic Risk Assessment: A Review. The University Of Sheffield, August, 2004. Disponível em: <http://www.shef.ac.uk > Acesso em: 15/01/2008.

DANESHKHAH, A. R; OAKLEY, J. & O'HAGAN, A. (2006). Nonparametric Prior Elicitation with Imprecisely assessed Probabilities. The University Of Sheffield, August, 2004. Disponível em: <http://www.shef.ac.uk > Acesso em: 15/01/2008.

DEVORE, Jay L. Probabilidade e Estatística: para engenharia e ciências. 6 ª edição. Tradução Joaquim Pinheiro Nunes da Silva. São Paulo: Pioneira Thomson Learning, 2006.

FLORES, Cláudio P.; NASSAR, Silvia; FREITAS FILHO, Paulo J. & MAGNO, Carlos. Risk Analysis using Monte Carlo simulation and Bayesian Networks. Proceedings of the 2006 Winter Simulation Conference. Monterey, CA, USA.

FREITAS FILHO, Paulo J. de. Introdução à Modelagem e Simulação de Sistemas. Florianópolis. Visual Books, 2001.

GARTHWAITE, P. H., KADANE, J. B. & O'HAGAN, A. Statistical methods for eliciting probability distributions. Journal of the American Statistical Association, 100, 680-701. 2005.

JANKAUSKAS, L. & MCLAFFERTY, S. Bestfit, Distribution Fitting Software by Palisade Corporation Proceedings of the 1995 Winter Simulation Conference. Arlington, VA, USA.

JENKINSON, D. The Elicitation of Probabilities – A Review of Statistical Literature. 72 p. April, 2005. Disponivel em:<www.shef.ac.uk> Acesso em: 15/01/2008.

KOKOSKA, S. & NEVISON, C. Statistical Tables and Formulae. USA. Springer-Verlag. 1989.

LARSON, Ron. & FARBER, Betsy. Estatística Aplicada. Tradução e revisão técnica Cyro de Carvalho Patarra. São Paulo. Prentice Hall, 2004.

LAW, A. M. & KELTON, W. D. Simulation Modeling and Analysis. 2º Edition. USA. McGraw-Hill, 1991.

LOVERIDGE, D. Experts and foresight: Review and experience. International Journal Foresight and Innovation Policy, v.1, n.1/2, p.33-69. 2004.

LUCENA, Bruno Rafael Dias de. Avaliação de recursos de petróleo não descobertos; metodologia e métodos de eliciação de informações subjetivas. Dissertação (mestrado) – Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Industrial, 2006.

MONTGOMERY, Douglas C. & RUNGER, George C. Estatística aplicada e probabilidade para engenheiros. 2ª Edição. Rio de Janeiro: LTC. 2003.

MOORE, D. S. & MCCABE, G. P. Introdução à Prática da Estatística. 3ª edição. Rio de Janeiro: LTC, 2002.

NETO, Pedro L. C. & CYMBALISTA, Melvin. Probabilidades: resumos teóricos, exercícios resolvidos, exercícios propostos. São Paulo. Editora Edgard Blucher Ltda. 1974.

NIST/SEMATECH. e-Handbook of Statistical Methods. Disponível em: http://www.itl.nist.gov/div898/handbook/index.htm. Acesso em: 20/03/2008.

OAKLEY, J. & O'HAGAN, A. Uncertainty in prior elicitations: a non-parametric approach. Department of Probability and Statistics, University of Sheffield, 2005. Disponível em:<www.shef.ac.uk> Acesso em: 15/01/2008.

O'HAGAN, A. & OAKLEY, J. E. Probability is perfect, but we can't elicit it perfectly. Reliability Engineering and System Safety, 85, 239-248. 2004.

O'HAGAN, A. Research in elicitation. Department of Probability and Statistics, University of Sheffield. Invited article for a volume entitled Bayesian Statistics and its Applications. 2005. Disponivel em:<www.shef.ac.uk> Acesso em: 15/01/2008.

ROMEU, J. L. The Chi-Square: a Large-Sample Goodness of Fit Test. START: Selected Topics in Assurance Related Technologies, Rome, v.10, n.4, p.1-6, 2003.

ROMEU, J. L. Anderson-Darling: A Goodness of Fit Test for Small Samples Assumptions. START: Selected Topics in Assurance Related Technologies, Rome, v.10, n.5, p.1-6, 2003.

ROMEU, J. L. Kolmogorov-Smirnov: A Goodness of Fit Test for Small Samples. START: Selected Topics in Assurance Related Technologies, Rome, v.10, n.6, p.1-6, 2003.

RUSSEL, Stuart J. & NORVIN, Peter. Inteligência Artificial. Tradução da segunda edição por PubliCare Consultoria. Rio de Janeiro: Elsevier, 2004.

SCOTT, Kendall. O processo unificado explicado. Tradução Ana M. Alencar Price. Porto Alegre. Editora Bookman, 2003

SILVA, Ricardo Pereira.UML2: Modelagem Orientada a Objetos. Florianópolis: Visual Books, 2007.

SPIEGEL, Murray R. Probabilidade e Estatística. São Paulo. McGraw-Hill do Brasil. Coleção Schaum, 1978.

TENÓRIO, Marcelo B. Reconhecimento de modelos de probabilidade. Dissertação (mestrado) - Universidade Federal de Santa Catarina, Departamento de Informática e Estatística, 2005.

THE FREE DICTIONARY. Disponível em: <http://www.thefreedictionary.com>. Acesso em: 20/03/2008.

WANG, H.; DAHS, D. & DRUZDEL, M.J. A Method for Evaluating Elicitation Schemes for Probabilistic Models. IEEE Transactions on Systems, Man, and Cybernetics-Part B:Cybernetics 32 38-43. 2002.